Two Novel Methods for Clustering Short Time-Course Gene Expression Profiles

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Division of Biomedical Engineering
University of Saskatchewan
Saskatoon

By
Wei Wei Fan

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Division of Biomedical Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5A9

Abstract

As genes with similar expression pattern are very likely having the same biological function, cluster analysis becomes an important tool to understand and predict gene functions from gene expression profiles. In many situations, each gene expression profile only contains a few data points. Directly applying traditional clustering algorithms to such short gene expression profiles does not yield satisfactory results. Developing clustering algorithms for short gene expression profiles is necessary.

In this thesis, two novel methods are developed for clustering short gene expression profiles. The first method, called the network-based clustering method, deals with the defect of short gene expression profiles by generating a gene co-expression network using conditional mutual information (CMI), which measures the non-linear relationship between two genes, as well as considering indirect gene relationships in the presence of other genes. The network-based clustering method consists of two steps. A gene co-expression network is firstly constructed from short gene expression profiles using a path consistency algorithm (PCA) based on the CMI between genes. Then, a gene functional module is identified in terms of cluster cohesiveness. The network-based clustering method is evaluated on 10 large scale *Arabidopsis thaliana* short time-course gene expression profile datasets in terms of gene ontology (GO) enrichment analysis, and compared with an existing method called Clustering with Overlapping Neighbourhood Expansion (ClusterONE). Gene functional modules identified by the network-based clustering method for 10 datasets returns target GO *p-values* as low as 10^{-24}, whereas the original ClusterONE yields insignificant results.

In order to more specifically cluster gene expression profiles, a second clustering method, namely the protein-protein interaction (PPI) integrated clustering method, is developed. It is designed for clustering short gene expression profiles by integrating gene expression profile patterns and curated PPI data. The method consists of the three following steps: (1) generate a number of predefined profile patterns according to the number of data points in the profiles and assign each gene to the predefined profile to which its expression profile is the

most similar; (2) integrate curated PPI data to refine the initial clustering result from (1); (3) combine the similar clusters from (2) to gradually reduce cluster numbers by a hierarchical clustering method. The PPI-integrated clustering method is evaluated on 10 large scale *A. thaliana* datasets using GO enrichment analysis, and by comparison with an existing method called Short Time-series Expression Miner (STEM). Target gene functional clusters identified by the PPI-integrated clustering method for 10 datasets returns GO *p-values* as low as 10^{-62}, whereas STEM returns GO *p-values* as low as 10^{-38}.

In addition to the method development, obtained clusters by two proposed methods are further analyzed to identify cross-talk genes under five stress conditions in root and shoot tissues. A list of potential abiotic stress tolerant genes are found.

Acknowledgements

I wish to express my deepest appreciation to my co-supervisors Professor Fang-Xiang Wu and Professor Gopalan Selvaraj for providing me with their guidance. This thesis would not have been completed without their help.

I would like to thank my committee members Professor Long-Hai Li and Professor Gordon Gray for offering me suggestions and comments on this thesis and also to thank my lab colleagues, in particular Bolin Chen for his clear and helpful explanation and teaching of many aspects of my work.

In addition, I thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Division of Biomedical Engineering in the College of Engineering at the University of Saskatchewan for supporting my work.

For my dearest parents, grandparents, boyfriend, friends, and to my own youth!

Is it possible that someday time could go backwards
Back to the leisurely years you and I can't regain
Perhaps. Someday. Even if the world ends
I still want to raise the memories — brewed sweet with you
To drink bottoms-up with you again.
If I had to choose a scene that represents youth
What comes to my mind are the tears and blue sky that year at graduation
Where we were crying while laughing and hugging.
Your faces fill me with love, longing, songs and tears
I miss those moments so
And longing always comes all of a sudden, with no forewarning.
When memories break free from test papers and burst past the years before my eyes
You and I, with sweat flowing, sip soft drinks beside the playground
Reach an agreement that we will go to the future world together.
Now, the world of the future is here
Why is it that your side and my side are no longer the same side?
Our vow of friendship was as strong as Noah's ark
But as I look out to sea, waiting for forever
My vision is forever blurred.

Is it possible that someday time could go backwards
Back to the leisurely years you and I can't regain
Perhaps. Someday. Even if the world ends
I still want to raise the memories — brewed sweet with you
To drink bottoms-up with you again.
Over the years I bought a car, a watch and a monocular
But I realize that what I cannot chase down and what I cannot stop are still the same

Life is simply accepting destiny and fate
All that is left makes it harder for us to laugh and easier to cry
But without making us get more mature
Maturity is a shattered fantasy, is training
Why is it that dreams grow smaller and smaller until they disappear?
Sometimes I feel like crying, but have no tears
Will you, will he, hold a reunion?
He is waiting for you, you are waiting for me, but who am I waiting for?
And whose children do not sleep, cell phone is uncharged, or mood is not unprepared?
The sky goes dark and light, then dark again
Times, places, and events fly past without pause
But we have no strength to chase them down anymore.

Is it possible that someday time could go backwards
Back to the leisurely years you and I can't regain
Perhaps. Someday. Even if the world ends
I still want to raise the memories — brewed sweet with you
To drink bottoms-up with you again.
Eventually, there will be a day
Where all of us become yesterdays
You have walked through life's journey alongside me
That day is today
Today is that day
I will tell you about all the gratitude I have been keeping inside
To drink bottoms-up with you again
One more drink for eternity
Drink so we will be able to live long
Years after years
Time has already stopped, and they already come back
Memorable people are waiting for my return (Ashin, 2011).

Contents

Permission to Use i

Abstract ii

Acknowledgements iv

Contents vii

List of Tables ix

List of Figures xi

List of Abbreviations xii

1 Introduction 1
- 1.1 Background . . . 1
  - 1.1.1 *Arabidopsis thaliana* data . . . 3
  - 1.1.2 Some effective clustering algorithms . . . 6
- 1.2 Objectives . . . 7
- 1.3 Organization of Thesis . . . 8

2 Data and result evaluation 10
- 2.1 Databases and Resources . . . 10
- 2.2 Datasets . . . 11
- 2.3 Gene Selection Method . . . 12
- 2.4 GO Enrichment Analysis . . . 14

3 Network-based clustering method 17
- 3.1 Overview . . . 17
- 3.2 Mutual Information . . . 20
- 3.3 Method . . . 21
  - 3.3.1 Gene co-expression network inference . . . 21
  - 3.3.2 Gene clustering with ClusterONE . . . 23
- 3.4 Result and Discussion . . . 27
  - 3.4.1 Result of method evaluation . . . 27
  - 3.4.2 Cross-stress comparison . . . 32
- 3.5 Conclusion . . . 35

4 PPI-integrated clustering method 36
- 4.1 Overview . . . 36
- 4.2 Method . . . 38
  - 4.2.1 Model profile pattern construction . . . 38

4.2.2 PPI incorporation to refine patterns 41
4.2.3 Hierarchical clustering for final patterns 43
4.3 Result and Discussion 44
4.3.1 Selection of significant profile patterns 44
4.3.2 Clustering results of *A. thaliana* datasets 45
4.3.3 Performance improvements through my proposed clustering method . 46
4.3.4 Comparison with STEM 47
4.3.5 Cross-stress comparison 57
4.4 Conclusion 60

5 Conclusions and Future Work **62**

References **67**

Appendix **78**

List of Tables

2.1 Number of abiotic-stressed *A. thaliana* microarray gene expression profiles analyzed in each database. . . . 11
2.2 The stress treatments and tissue information of 10 *A. thaliana* microarray gene expression datasets measure in Affymetrix array with samples grow at the seedling stage. . . . 12
2.3 Number of input genes for cluster analysis after gene selection for ten *A. thaliana* datasets downloaded from GEO. . . . 13

3.1 Thresholds of edge deletion for MI and CMI in the process of generating gene co-expression networks for ten datasets. . . . 28
3.2 The number of annotated genes assigned to stress-related GO terms in *A. thaliana* for seed selection. . . . 28
3.3 Clustering results of *A. thaliana* root tissue datasets from the network-based clustering method. . . . 29
3.4 Clustering results of *A. thaliana* shoot tissue datasets from the network-based clustering method. . . . 30
3.5 Clustering results of *A. thaliana* root tissue datasets from ClusterONE. . . . 30
3.6 Clustering results of *A. thaliana* shoot tissue datasets from ClusterONE. . . . 31
3.7 A list of cross-talk genes under five abiotic stresses from *A. thaliana* root tissue at the seedling stage by network-based approach. . . . 32
3.8 A list of cross-talk genes under five abiotic stresses from *A. thaliana* shoot tissue at the seedling stage by network-based approach. . . . 33

4.1 Numbers of significant model profile patterns after PPI refinement out of total constructed model profile patterns for 10 *A. thaliana* datasets. . . . 44
4.2 Clustering results of *A. thaliana* root tissue datasets from the PPI-integrated clustering method. . . . 48
4.3 Clustering results of *A. thaliana* shoot tissue datasets by the PPI-integrated clustering method. . . . 48
4.4 Clustering results of *A. thaliana* root tissue datasets by STEM. . . . 49
4.5 Clustering results of *A. thaliana* shoot tissue datasets by STEM. . . . 49
4.6 Comparison of stress-associated molecular functions and biological processes in obtained clusters from datasets generated from *A. thaliana* root tissue samples by my proposed method and STEM. . . . 55
4.7 Comparison of stress-associated molecular functions and biological processes in obtained clusters from datasets generated from *A. thaliana* shoot tissue samples by my proposed method and STEM. . . . 56
4.8 A list of cross-talk genes under five abiotic stresses from *A. thaliana* root tissue at the seedling stage by PPI-integrated clustering approach. . . . 57
4.9 A list of cross-talk genes under five abiotic stresses from *A. thaliana* shoot tissue at the seedling stage by PPI-integrated clustering approach. . . . 58

5.1 Overlap scores of stress-responsive clusters produced by network-based and PPI-integrated clustering methods of ten datasets. 63

List of Figures

1.1 Intra-cluster and Inter-cluster relations. 4
1.2 General workflow of my study, where steps mark in red indicate the contributions of my work. 9

2.1 Advantage of the logarithms transformation. 14

3.1 Schematic of the network-based clustering method. 18
3.2 Illustration of how sub-network is generated from the seed gene. 27

4.1 Schematic of the PPI-integrated clustering method. 38
4.2 Clustering results for ten *A. thaliana* datasets. 45
4.3 Reduction of GO *p-values* from step 1 to step 3 of the PPI-integrated clustering method for five datasets generated from the *A. thaliana* root tissue. 46
4.4 Reduction of GO *p-values* from step 1 to step 3 of the PPI-integrated clustering method for five datasets generated from the *A. thaliana* shoot tissue. 46
4.5 Example of cold- and drought-induced genes functions in *A. thaliana*. 52
4.6 A portion of directed acyclic graph of GO term domain relations. Term box (child term) at the root of arrow belongs to term box (parent term) at the tip of arrow. 53

List of Abbreviations

ClusterONE	Clustering with Overlapping Neighbourhood Expansion
CMC	Clustering based on Maximal Cliques
CMI	Conditional Mutual Information
DNA	DeoxyriboNucleic Acid
GEO	Gene Expression Omnibus
GO	Gene Ontology
MCL	Markov Cluster
PCA	Path Consistency Algorithm
PCC	Pearson Correlation Coefficient
PCR	Polymerase Chain Reaction
PLEXdb	Plant Expression Database
PPI	Protein-Protein Interaction
qPCR	real-time Polymerase Chain Reaction
RNA	RiboNucleic Acid
RNSC	Restricted Neighbourhood Search Clustering algorithm
STEM	Short Time-series Expression Miner
TAIR	The Arabidopsis Information Resource

Chapter 1

Introduction

1.1 Background

A gene is a molecular unit that carries genetic information in living organisms. It is almost always comprised of DNA (deoxyribonucleic acid) but in some viruses it is of RNA (ribonucleic acid). In *Arabidopsis thaliana*, a model plant, the genome is composed of around 135 million DNA base pairs, comprising 25,498 genes (The Arabidopsis Genome Initiative, 2000). With the sequence in hand, genetic variants in the genome sequence can be identified, which in some cases increase the risk of cancer or cause inheritable diseases. Genetic information from DNA is transcribed into messenger RNA (mRNA) and then translated into proteins, which perform various functions in the organism. The process by which the gene information is used in the synthesis of functional gene products is collectively known as gene expression. Generally, genes are expressed constitutively or in response to some stimuli through the molecular process of gene regulation, which acts as a switch to turn genes on and off to allow cells to express proteins when needed. For example, when a person suffers a cut of his or her skin, the tissue will respond to the wound stimuli by activating a healing process. This process involves regulation of gene expression for programmed replacement of old cells with new cells. The skin tissue sample can be taken from the person and the gene expression levels can be monitored to determine the genes that are responsible for skin cut defense.

Generally, when plant species are not growing in ideal environmental conditions, plants are considered to be under stress. Such stress conditions will prevent plants from reaching their maximum growth, development and productivity. Stress consists of biotic stress and abiotic stress, in which biotic stress is the negative impact of living organisms on plants, such as

bacteria, fungi, viruses and so on. Similarly, abiotic stress is the negative impact of non-living factors on plants, such as extreme temperatures, drought, osmotic gradient, salinity and so on. Both biotic stress and abiotic stress can reduce plant productivity by 65% to 87% (Buchanan *et al.*, 2002). In plant biology, stress response is the response of plants to an environmental condition or a stimulus. It is the method by what a plant reacts to an external challenge. A stress response will be initiated to activate signal transduction pathways when plants recognize stress at the cellular level. As a result, there will be changes in gene expression levels to influence the reproductive capacities of plants. Therefore, understanding the activity of genes involved in plant stress responses can bring up plant productivity. The use of microarray for gene expression profiling is capable for measuring the activities of thousands of genes at a time, which can create a global picture of genes' cellular functions.

Recently, many technological advances have enabled large-scale gene expression studies. DNA microarray technology is one example of gene expression profiling that is designed for quantifying DNA gene expression on a large scale, such as the whole genome level. In order to accurately perform the microarray experiment, a particular guideline needs to be followed to make sure experiments are properly conducted, which comes to the introduction of MIAME (minimum information about a microarray experiment) guideline for the microarray experiment. It describes the minimum information about a microarray experiment that is required to unambiguously interpret the results produced from the experiment, as well as to allow the experiment being replicated by other researchers (Brazma et al., 2001). Six important elements must be provided to support microarray based work. They are the raw data files produced by the microarray imaging analysis softwares, the processed data after normalization, the experimental factors and their corresponding values, the experimental design description, annotation of the array design and the experimental and data processing protocols.

Such large-scale data produced from microarray experiments can be used to predict gene functions. For the prediction of gene functions, gene expression levels need to be measured at different time points during a biological process of interest. From such data, gene expres-

sion patterns can be recognized and analyzed. Genes, whose expressions follow a specific pattern, are clustered.

1.1.1 *Arabidopsis thaliana* data

The gene expression data of the flowering plant *A. thaliana* can be used for cluster analysis, as *A. thaliana* offers advantages for computation and biological research to validate computational models, as listed below:

(1) Small genome size

(2) Fully sequenced genome

(3) Available mutants for future research

(4) Available well-annotated protein-protein interaction information

(5) Short generation time

(6) Large amount of offspring

(7) Easy to treat with various stresses

Generally, clustering can be used to help understand microarray data. In bioinformatics, clustering is used for predicting gene functions from high-throughput data, as well as for understanding gene regulatory pathways (Eisen *et al.*, 1998; Tavazoie *et al.*, 1999). Clustering is an important step in associating novel genes to predicted functional patterns.

The concept of data cluster analysis can be formulated as follows: given a set of n objects, denoted by x_1, x_2, x_3 ... x_{n}, the task of clustering is to divide the objects into a number of groups so that objects in the same group are more similar than those in different groups. Typically each object can be represented by m numerical values. As a result, all n objects

can be expressed in the following n by m matrix:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ ... & ... & ... & ... & ... \\ x_{n1} & x_{n2} & x_{n3} & ... & x_{nm} \end{bmatrix}$$

In bioinformatics, cluster analysis is a helpful exploratory technique to group n genes with similar expression patterns x_1, x_2, x_3 ... x_n (co-expressed genes) into the same cluster, where genes in the same cluster are more similar to each other in biological function than genes in different clusters. The goal of cluster analysis is to obtain high intra-cluster similarity, but low inter-cluster similarity as illustrated in Figure 1.1.

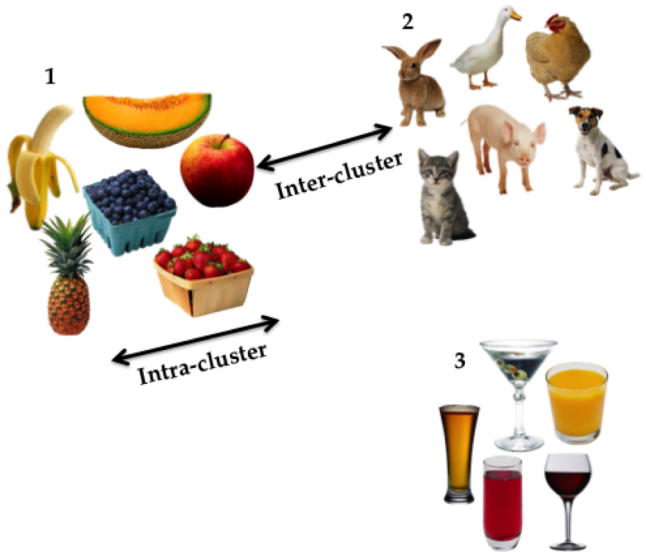

Figure 1.1: Intra-cluster and Inter-cluster relations.

Many gene-clustering algorithms are developed by employing different cluster models (Fraley and Raftery, 1998); such as connectivity models, distribution models, centroid models, density models, graph-based models, group models and subspace models. The key point in understanding how different clustering algorithms interpret data is to understand the various types of cluster models algorithms used to define a cluster (Boja, 2011).

For example, the well known "k-means clustering" (Arai and Barakbah, 2007) method uses a centroid model, which specifies a mean vector that is not necessarily an object of the dataset

to represent a cluster. Objects are assigned to the closest predefined clusters based on the distance calculation between objects and clusters. The clustering algorithms based on the centroid model are examples of partitioning hard clustering algorithms, by which objects in a dataset are divided into a number of distinct clusters such that each object belongs to exactly one cluster. The drawbacks of clustering algorithms based on the centroid model are that the number of clusters in a given dataset needs to be predefined, and it is highly sensitive to the data noise and data outliers. Therefore, they only work well for dataset clusters with similar sizes and dataset clusters with high intra-cluster similarities.

"Hierarchical clustering" (Joe, 1963) is an example of a connectivity model that builds up clusters based on distance connectivity. The similarity relationship of objects in a dataset can be represented using a dendrogram. The linkage criterion of hierarchical clustering determines how objects in a dendrogram are connected, thus different choices of linkage criterion lead to different dendrogram structures. The single linkage criterion merges objects by measuring the minimum of object distance, where as the average linkage criteria merges data objects by measuring the average of object distance, and so on. The connectivity-based clustering algorithms do not require a predefined number of clusters. However, these algorithms tend to have at least quadratic complexities that are relatively time consuming compared to other model-based clustering algorithms. The running time of such connectivity model based clustering algorithms is highly dependent on the number of objects in the dataset.

The effectiveness of gene cluster analysis is highly dependent on two factors: the selection of a similarity function for measuring gene relationships and the effect of data noise. The similarity function for gene relationships defines the relationship between genes in a set, becoming the fundamental rule by which genes are clustered. Selecting a similarity function that is ideal for the application can significantly improve clustering effectiveness. The most commonly used method for measuring gene relationships for gene expression data is standard correlation measure, such as Pearson correlation coefficient (PCC) method, which measures the linear relationship between gene profiles. In addition, the nonlinear relationships between gene profiles as the generalized correlation measurement, such as mutual information (MI)

can be assessed.

No clear evidence has been found to show which similarity function outperforms the others in studies with different purposes, as many articles applied correlation coefficient as gene co-expression measure (Eisema *et al.*, 1998; Zhou *et al.*, 2002; Stuart *et al.*, 2003; Zhang and Horvath, 2005; Langfelder and Horvath, 2008), while others applied MI as gene co-expression measure (Butte *et al.*, 2000; Daub *et al.*, 2004; Basso *et al.*, 2005; Margolin *et al.*, 2006; Meyer *et al.*, 2008; Cadeiras *et al.*, 2010). The application of the MI method can significant reduce erroneous clustering results possibly generated from data noise, which is a common problem for cluster analysis (Priness *et al.*, 2007). Therefore, either integrating other types of data for better data coverage, or developing ideal algorithms to reduce the effects of noise can overcome the effect of data noise.

No perfect algorithm exists, due to the reason that cluster analysis is an iterative process of knowledge discovery that involves trials and failures. Furthermore, in many situations, each gene expression profile in microarray data contains few data points due to either the high cost of microarray experiments, or the limited amount of genetic material.

1.1.2 Some effective clustering algorithms

A. Short Time-Series Expression Mining (STEM) Algorithm

Ernst et al. (2005) developed the STEM algorithm by employing the centroid-clustering model to categorize genes. The algorithm has advantages in effectively grouping short time-course microarray gene expression data. Because quantitive gene expression patterns are used for gene clustering analysis, genes with few time points (3 to 8 time points) will produce statistically insignificant results due to large differences between the number of time points and the number of genes available for clustering (more than 10,000). The STEM algorithm overcomes this problem by considering the expression pattern that occurs between each consecutive time point, which increases the number of possible pattern combinations that form the mean vectors, which represent clusters, rather than the expression pattern of the whole

time-series.

Algorithm effectiveness is highly dependent on the input data. The disadvantage of the STEM algorithm is that it uses microarray data as the only data source for cluster analysis, which might produce inaccurate results if the data contains high levels of background noise, which is common for microarray data.

B. ClusterONE (Clustering with Overlapping Neighbourhood Expansion) Algorithm

Nepusz et al. (2012) introduced a clustering method for identifying protein complexes in the protein-protein interaction (PPI) network, where substructures with high degrees of internal connectivity but low degrees of connectivity to the rest of the nodes in the network are thought to be protein complexes. The ClusterONE algorithm currently is the most efficient algorithm for identification of protein complexes in the PPI network, in comparison to other clustering approaches including MCL (Markov Cluster), CMC (Clustering based on Maximal Cliques), RNSC (Restricted Neighbourhood Search Clustering algorithm) (Stijn, V. D., 2000; Liu *et al.*, 2009; Andrew, 2004). For the clustering of gene expression profiles, the logic follows the same concept as the identification of protein complexes, where genes within the same cluster are more closely related in distance than to the rest of genes in other clusters. The application of ClusterONE analysis to identifying protein complexes does help to identify gene clusters in gene co-expression networks, but it has limitations with respect to cluster sizes, as gene cluster sizes are usually large, whereas the number of proteins in a protein complex is relatively small. ClusterONE needs to be improved before it is applicable to gene expression data.

1.2 Objectives

The objective of this thesis is to develop methods for clustering short gene expression profiles based on the selection of the similarity function for gene relationships and the reduction of data noise, with the goal of improving the effectiveness of clustering results. Specifically, the

following issues are addressed:

(1) To infer a gene co-expression network from short time-course gene expression profiles based on conditional mutual information (CMI), and then to identify modules/clusters of interested from the inferred gene co-expression network in terms of gene ontology (GO) enrichment analysis.

(2) To improve clustering results in terms of the GO *p-value* from short microarray gene expression profiles by integrating PPI data for better data coverage, and thus to more effectively categorize genes into functional clusters.

(3) To identify cross-talk genes among various abiotic stress conditions by cross-comparing stress-responsive clusters obtained using the proposed methods, thus to identify a list of potential abiotic stress tolerant genes.

1.3 Organization of Thesis

In Chapter 2, the background information required to complete the method development and data evaluation work of this thesis is presented. It includes the introduction of various databases used to collect the *A. thaliana* datasets on which the proposed methods are tested. Chapter 2 introduces detailed information about the *A. thaliana* datasets, including information of the experimental design and data pre-processing method. Lastly, Chapter 2 introduces the methods for evaluation of results, in which the clustering methods developed here are compared to those currently in use.

In Chapter 3, the proposed network-based clustering method is introduced. The concept of mutual information is explained as the selected similarity function for measuring the gene relationship. Next, the path consistency algorithm for inferring a gene co-expression network based on conditional mutual information (CMI) from short microarray gene expression profiles is explained. Third, a method for identifying the stress-responsive functional module

from the generated gene co-expression network is introduced. Finally, the results are presented with a discussion of the network-based clustering method performance in terms of GO enrichment analysis.

In Chapter 4, a PPI-integrated clustering method is presented. Some background information on the requirements needed to develop this PPI-integrated clustering method is presented. Then, the procedures of the method are described in three steps. Finally, the clustering result is presented accompanied by the evaluation of the method by GO enrichment analysis.

In Chapter 5, the results are summarized and the conclusions are drawn, followed by suggested directions for further researches.

The general workflow for cluster analysis, and the summarized workflow of my study is shown in the following figures. The cluster analysis and cross-talk genes boxed in red are the main contributions of this thesis.

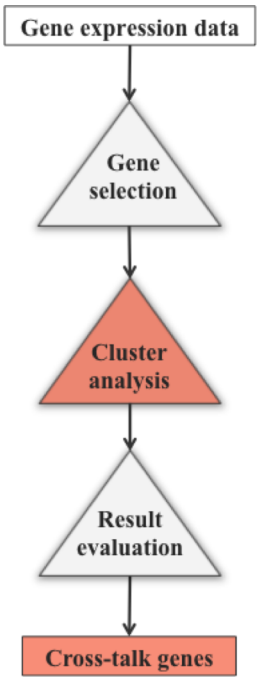

Figure 1.2: General workflow of my study, where steps mark in red indicate the contributions of my work.

Chapter 2

Data and result evaluation

In this chapter, the databases used for this thesis are introduced. Databases are used for either collecting datasets for method evaluation, or for evaluating the clustering results by proposed methods. Then, the data pre-processing procedures, including data logarithm transformation and data normalization, are described. Later, the method for gene selection prior to cluster analysis and method for result evaluation posterior to cluster analysis are introduced. Finally, references are provided for previously developed algorithms with which my methods were compared.

2.1 Databases and Resources

Two types of databases are used, namely databases for data collecting and for result evaluation:

Databases analyzed for data collecting were GEO (Gene Expression Omnibus), TAIR (The Arabidopsis Information Resource), PLEXdb (Plant Expression Database) and ArrayExpress (Edgar *et al.*, 2002; Swarbreck *et al.*, 2008; Dash *et al.*, 2012; Parkinson *et al.*, 2003).

Tool for result evaluation was GO::TermFinder tool (Boyle *et al.*, 2004). This tool comprises a set of Perl modules (components of software for the Perl programming language), which is able to access GO information to evaluate and visualize the GO annotation of a list of input genes to GO terms. The tool is able to access GO annotation files from various information resources, and download the most up-to-date GO annotation file for any specified species. For my thesis work, GO::TermFinder gets the *A. thaliana* GO annotation file

from TAIR, which is the most recent version, as the ontology and gene association files are downloaded nightly from information resources.

In order to collect the datasets for this study, the biological databases mentioned were explored and all the suitable datasets for the application to this thesis were gathered. The collected microarray datasets should contain different stressed samples from different tissues, to ensure cross-stress comparison in various tissue samples could be performed. The number of *A. thaliana* microarray gene expression profiles in each database for this thesis are summarized in Table 2.1.

Table 2.1: Number of abiotic-stressed *A. thaliana* microarray gene expression profiles analyzed in each database.

GEO	**Plexdb**	**ArrayExpress**	**TAIR**
10	10	7	6

The 10 datasets from GEO and Plexdb used exactly the same experimental design with the Affymetrix platform. ArrayExpress contains 7 datasets and TAIR contains 6 datasets with microarray platforms of Affymetrix and Carnegie, respectively.

2.2 Datasets

Ten datasets from GEO (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37118) were used for method evaluation, as the datasets comprise 5 different types of abiotic stresses (**cold, drought, heat, salt stress and osmotic stress**) from 2 types of sample tissues (**root and shoot**) at the seedling stage of *A. thaliana*. The 10 microarray gene expression profiles were collected from one lab, and datasets had been used for publication (Heinrich *et al.*, 2012). A variety of stress types and tissue types enable a cross-stress comparison of clustering results useful for identification of commonly expressed genes, especially for identification of commonly expressed unannotated genes.

Each dataset contains time series expression levels of 22721 genes. Each gene has seven time points (0, 0.5, 1, 3, 6, 12, 24hrs) with two replications for each time point by Affymetrix ATH1 Arabidopsis Genome Array. I decided the value for each time point based on the agreement between two replications. The integrated *A. thaliana* PPI data was the curated PPI data, which was downloaded from TAIR database (www.arabidopsis.org). All interactions are derived from literature curation based on the binary interaction of proteins in *A. thaliana*.

The detailed information of the ten datasets is listed in Table 2.2.

Table 2.2: The stress treatments and tissue information of 10 *A. thaliana* microarray gene expression datasets measure in Affymetrix array with samples grow at the seedling stage.

	Treatment	**Tissue**
1	Cold	Root
2	Drought	Root
3	Heat	Root
4	Salinity	Root
5	Osmotic stress	Root
6	Cold	Shoot
7	Drought	Shoot
8	Heat	Shoot
9	Salinity	Shoot
10	Osmotic stress	Shoot

2.3 Gene Selection Method

Gene expression values are measured in terms of intensity in microarray experiment. Gene intensity values can have a large numerical range from zero up to several thousands. There-

fore, to plot a gene distribution curve, which shows the number of genes in different intensity values, would generate an extreme skewed distribution curve. In order to avoid that, intensity values are Log_2 transformed in order to narrow down the data range for better presentation and easier comparison. That logarithms transformation can significantly reduce data skewness as shown in Figures 2.1, where the top figure illustrates the distribution of expression data before log transformation, and the bottom figure shows the distribution of expression data after log transformation.

After the microarray gene expression intensities are Log_2 transformed, the Log_2 transformed expression values are shifted, so that the time series starts at 0. The fold change for each time point of a gene is calculated as follows:

X = measured value (stress-treated value); Y = control value (control value)

$$Foldchange = Log_2X - Log_2Y \tag{2.1}$$

Prior to cluster analysis, differentially expressed genes with at least one maximum absolute expression value greater than two-fold were inputted for cluster analysis (Dalman *et al.*, 2012). The fold-change value measures the relative expression of a gene in a treatment condition with respect to the control. The numbers of selected genes from each dataset for the input of cluster analysis are presented in Table 2.3.

Table 2.3: Number of input genes for cluster analysis after gene selection for ten *A. thaliana* datasets downloaded from GEO.

Plant tissue	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
Root	8,444	5,658	8,592	10,692	8,819
Shoot	11,445	6,851	9,276	8,171	11,787

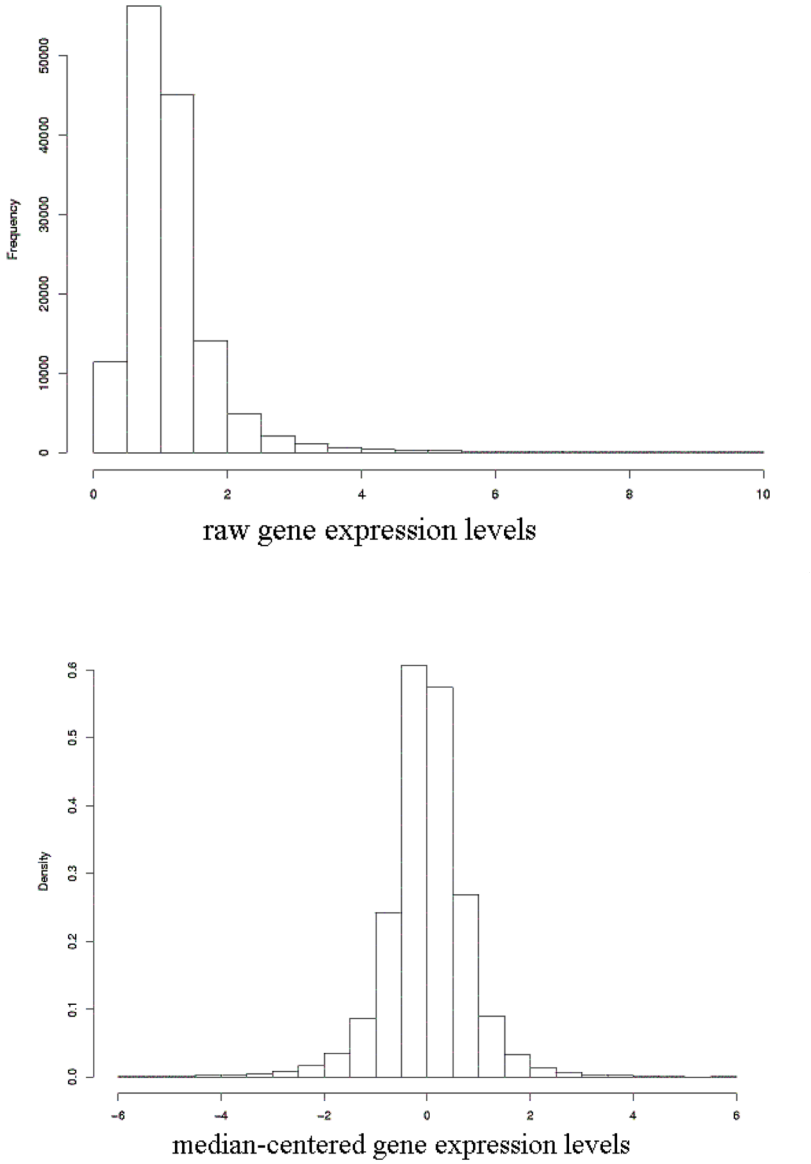

Figure 2.1: Advantage of the logarithms transformation.

2.4 GO Enrichment Analysis

Following cluster analysis, results generated from the proposed clustering methods must be evaluated in order to assess the effectiveness of the methods. GO enrichment analysis was used to generate the final clusters used in this thesis.

GO is the most popular database, aiming at specifying consistent representative terminologies of genes and gene products across different species and databases. In this database, GO terms are structured in a directed acyclic graph, with the terms being categorized in three ontology domains, including cellular components, biological processes and molecular functions. Due to the reliability of GO database that is summarized and addressed from the literature and by professionals, most researchers use GO database to assess gene-clustering results.

The most common approach using GO database is to search for gene annotations of a list of genes, and determine whether the observed number of annotations for the list of genes is significantly enriched within the context of annotations for all genes in the genome. GO::TermFinder is a tool for gene enrichment analysis (Boyle *et al.*, 2004). The tool comprises a set of Perl modules to access GO information to evaluate and visualize the GO annotation of a list of input genes to GO terms. The tool is able to access GO annotation files from various information resources for downloading the most up-to-date GO annotation file for specified species. For this work, GO::TermFinder gets the *A. thaliana* GO annotation file from TAIR, which is the most up-to-date version, as the ontology and gene association files are downloaded nightly from information resources. GO:TermFinder categorizes genes according to GO terms for a list of genes, and returns GO terms with significance level represents by a *p-value* below the threshold along with their assigned genes. The GO *p-value* is a measurement of significance level, which represents the probability that the observed numbers of counts resulted from random distributing the GO term between tested gene group and reference gene group using Equation 2.2. It determines whether the GO term annotates a list of genes at a frequency greater than expected by chance. The GO *p-value* can be calculated according to hyper-geometric distribution as follows (Boyle *et al.*, 2004).

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}, \tag{2.2}$$

where N is the number of total genes in the background distribution, M is the number of genes that are annotated to the GO term in the background distribution, n is the number of

genes in the obtained stress-related cluster, k is the number of genes in the obtained stress-related cluster that are annotated to the GO term. The background distribution is all genes in the given annotation file, which in this work is all annotated genes from *A. thaliana*.

However, in case that more than one statistical test is performed, the chance of finding at least one significant annotation by chance will increase. Therefore, the Bonferroni correction is used to adjust the significance level of an individual test to the predefined threshold. The lower Bonferroni corrected *p-value* is, the more significant are the genes with the annotation. Annotations are represented in the name of GO term. In this work, the following clustering of genes under stress conditions are: GO term: **response to cold (GO:0009409)** for the dataset generated from cold treated samples, GO term: **response to water deprivation (GO:0009414)** for the data generated from drought treated samples, GO term: **response to heat (GO:0009408)** for the data generated from heat treated samples, GO term: **response to salt stress (GO:0009651)** for the data generated from salt treated samples and GO term: **response to osmotic stress (GO:0006970)** for the data generated from osmotic treated samples. The corrected *p-values* of above listed GO terms were used for evaluating the method performance.

The proposed gene clustering methods for short time-course gene expression profiles are mainly evaluated by comparing with the previously developed STEM (Ernst *et al.*, 2005) and ClusterONE (Nepusz *et al.*, 2012) algorithms.

Chapter 3

Network-based clustering method

3.1 Overview

Most of the current clustering algorithms are suitable for clustering gene expression profiles consisting of 10 or more time points. This is because the clustering algorithms are based upon a pairwise coupling of the most similar gene variables of long time-course gene expression profiles to produce statistically significant results (Kavitha and Punithavalli, 2010). However, more than 80% of gene expression profiles produced from microarray technology contain fewer time points per gene, which are often very sensitive to small changes between times (Ernst *et al.*, 2005). Therefore, to develop a clustering method, which can effectively cluster short time-course gene expression profiles, would have a major impact on the analysis of microarray data.

Short time-course gene expression profiles for cluster analysis often have technical issues and experimental noise issues, originating from the probe hybridization, signal variations, different experimental handling and low replications (Tu *et al.*, 2002; Ricardo and Tie, 2005). Gene expression profiles are snapshots of a dynamic system (network) in which genes regulate each other through their products (proteins or RNAs). Therefore, to better understand the regulatory relationships, a network-based clustering method which can effectively discover modular structures in short time-course microarray gene expression profiles, was developed in this thesis.

The proposed network-based clustering method consists of three parts: 1) A gene co-expression network is firstly constructed from short time-course gene expression profiles using path con-

sistency algorithm (PCA), based on conditional mutual information (CMI) between gene variables, 2) a sub-network is grown from the selected seed from the inferred gene co-expression network based on "second-order neighbouring search" method, and 3) A gene functional module is obtained from the sub-network by iteratively calculating the cluster cohesiveness score. The workflow of the network-based clustering method is illustrated in figure 3.1 within the dashed line box.

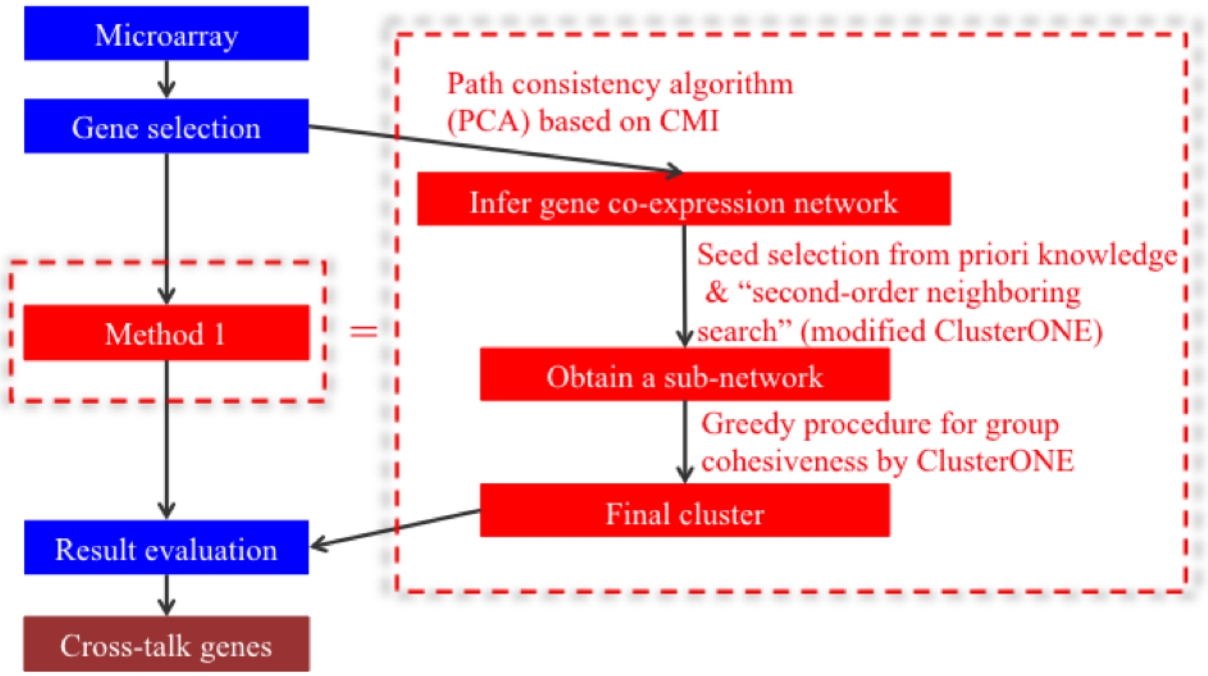

Figure 3.1: Schematic of the network-based clustering method.

The first part of the network-based clustering method is associated with gene co-expression network inferences. Generally speaking, a network consists of a set of nodes is connected by either direct or indirect edges. A biological network inference is a process of making predictions about biological networks with predefined relationships connected by the edges in the network. The concept of gene co-expression network inference is straightforward as nodes represent genes, and nodes are connected if two gene variables are positively co-expressed according to predefined threshold. Such a gene co-expression network is defined as the unweighted gene co-expression network, because the binary information (above threshold: connected = 1, below threshold: unconnected = 0) is used for encoding the gene co-expression

network. The purpose is to extract as much biological information as possible from the short time-course microarray gene expression data, in order to predict a gene co-expression network that reliably resembles the co-expression relationship of genes in the cell environment. Current algorithms construct gene co-expression network using PCC, which is based solely on the linear relationships between genes (Keiichi *et al.*, 2011). In this study, the gene co-expression relationships are measured using CMI, which takes into accounts the nonlinear relationship between any two connected genes in the presence of other genes. The nonlinear measurement of gene relationship is considered to be a more accurate method than linear measurement, as gene profiles are not always linearly proportional to one another (Pele *et al.*, 2013). Therefore, CMI is employed to compute the interdependence between gene pairs, in the presence of a third gene. The concept of this network-based clustering method takes the neighbouring genes of each gene-pair into consideration, which makes use of more information in the available data for better prediction of gene relationships in the network. The proposed method infers gene co-expression network from short gene expression data by using path consistency algorithm (PCA) based on CMI, which was originally developed for the inference of gene regulatory network from (Zhang *et al.*, 2012).

The second part of the network-based clustering method is to grow the sub-network from the selected seed in the inferred gene co-expression network based on "second-order neighbouring search" method. The method initially selects a seed gene, which has the highest internal connectivity among a list of known stress-responsive genes to grow the "sub-network", whose nodes are expected to be closely related to stress. The procedure to obtain the sub-network starts from the expansion of selected seed gene will be introduced in details in the section 3.3.

The third part of the network-based clustering method is to perform a greedy procedure to calculate the group cohesiveness score, in order to identify a dense-cluster from the sub-network as the final cluster. This is assuming that a densely connected structure from the obtained sub-network is more likely to form a cluster of functionally related genes.

3.2 Mutual Information

Mutual Information (MI) is a measurement of the mutual dependence between two random variables. CMI is the mutual dependence between two random variables given a third conditional variable. MI and CMI can be computed using concise formulas that involve covariance matrices of the corresponding gene expression profiles (Zhang et al, 2012). MI has been used in many reports to construct gene co-expression networks from microarray gene expression data (Altay and Emmert, 2010). MI is based upon the following. Let the expression of gene 1 represented by random variable x and the expression of gene 2 represented by random variable y, where gene expression profiles with n time points can be viewed as the sample data from these random variables, as illustrated below:

Let $x_i = [x_{\mathrm{i1}}, x_{\mathrm{i2}}, x_{\mathrm{i3}} \ldots x_{\mathrm{in}}]$, therefore, $\bar{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{\mathrm{ij}}$
The covariance of genes in Equations 3.2, 3.3 and 3.4 can be calculated as the formula below:

$$C(x_1, x_2, x_3 \ldots x_{\mathrm{m}}) = \frac{1}{n-1} \begin{bmatrix} x_1 - \bar{x_1} \\ x_2 - \bar{x_2} \\ x_3 - \bar{x_3} \\ \cdots \\ x_m - \bar{x_m} \end{bmatrix} \times \begin{bmatrix} x'_1 - \bar{x_1} & x'_2 - \bar{x_2} & x'_3 - \bar{x_3} & \ldots & x'_m - \bar{x_m} \end{bmatrix} \tag{3.1}$$

Therefore, the MI of two continues random variables x and y can be calculated by the formula below (Zhang *et al.*, 2012):

$$\textbf{MI:} \qquad I(x,y) = \frac{1}{2} Log \frac{|C(x)| * |C(y)|}{|C(x,y)|}, \tag{3.2}$$

where C is the variance and covariance matrix of the defined variables and $\mid C \mid$ is the determinant of matrix C. As a result, the degree measurement of dependence between genes allows the inference of gene co-expression network, which shows how closely two genes are related to each other. If two independent genes are regulated by the same set of genes in the network, their expression profiles are very similar. Based on the MI of their profiles, one can

conclude that these two genes are closely related to each other. To help exclude this kind of false positives, CMI is adopted to measure the dependence between two genes given their commonly regulatory gene profiles, where CMI between variables x and y given variable z can be calculated as follow (Zhang *et al.*, 2012):

$$\textbf{CMI:} \qquad I(x,y|z) = \frac{1}{2} Log \frac{|C(x,z)| * |C(y,z)|}{|C(z)| * |C(x,y,z)|}, \tag{3.3}$$

where C is the variance/covariance matrix of the defined variables and $\mid C \mid$ is the determinant of matrix C. Equations 3.2 and 3.3 are used in PCA for calculating dependence between genes for generating the gene co-expression network.

For higher order CMI with more than one variable conditions, the formula below is used, where conditional gene variables are denoted as $w, v, ...$

$$\textbf{CMI:} \qquad I(x,y|w,v,...) = \frac{1}{2} Log \frac{|C(x,w,v,...)| * |C(y,w,v,...)|}{|C(w,v,...)| * |C(x,y,w,v,...)|} \tag{3.4}$$

3.3 Method

3.3.1 Gene co-expression network inference

The inference of a gene co-expression network starts from a complete network, and then the edges are removed between any two genes with independent relationship. (If the dependence value between two genes is below the pre-defined threshold, two genes have independent relationship). The independence value between genes is calculated based on either the MI or CMI values. The process of edge removal is carried out by PCA as follows (Zhang et al, 2012).

First, a complete gene co-expression network is generated by connecting all possible gene-pairs among a list of differentially expressed genes, denoted as G. MIs are calculated for all

possible combinations of gene-pairs in G. A threshold is predefined to delete any edges that connect two genes with a MI value below the threshold, in which the deleted-edges that connect gene $m(m \in G)$ and gene $n(n \in G)$ indicate independence relationships between these two genes. The obtained network after the first round of edge deletion based on $MI(m, n)$ is referred to as the zero-order network.

The next step is to generate the first-order network by computing CMI between any two connected genes (m and n) in the zero-order network, given a list of their parent genes, denoted as l_{i}, where $i = 1....k$ (k = the number of their parent genes). Here, parent genes are a set of genes regulating both gene m and gene n from the generated zero-order network, which connect to both m and n. The notion for calculating CMI to generate the first-order network can be represented by $CMI(m, n|l_{\mathrm{i}})$. For each gene-pair (m, n), if one $CMI(m, n|l_{\mathrm{i}})$ $(i = 1, \ldots, k)$ value is below the predefined threshold, edges between m and n are deleted. As the selection order for the gene-pair matters, the algorithm by default selects a gene-pair with the highest number of parent genes to first calculate $CMI(m, n|l_{\mathrm{i}})$. After $CMI(m, n|l_{\mathrm{i}})$ is calculated for the selected gene-pair, and decision has been made (to delete the edge or keep the edge), network is updated to calculate $CMI(m, n|l_{\mathrm{i}})$ for another pair of connected genes, until all connected gene-pairs have been considered. The obtained gene co-expression network after the second round of edge deletion based on $CMI(m, n|l_{\mathrm{i}})$ values is referred to the first-order network.

After obtaining the first-order network, the algorithm continues to generate the second-order network by computing $CMI(m, n|l_{\mathrm{i}}, l_{\mathrm{j}})$, in which l is a list of parent genes for m and n $(i, j = 1, \ldots, k$, k = the number of parent genes) between any two edge-connected genes (m and n) in the first-order network. In this case, the CMI of any two edge-connected genes is computed, given any two parent genes from l. For each gene-pair (m, n), if there is one value of $CMI(m, n|l_{\mathrm{i}}, l_{\mathrm{j}})$ below the predefined threshold, edges between m and n are deleted. Network is refreshed to calculate CMI for another pair of edge-connected genes, until all edge-connected gene-pairs have been tested. The obtained gene co-expression network after the third round of edge deletion based on $CMI(m, n|l_{\mathrm{i}}, l_{\mathrm{j}})$ values is called the second-order

network.

By default, the algorithm should continue to generate higher-order networks, until no further edges can be deleted from the network. However, the inference of gene co-expression network only goes up to the fourth order, since the applied differentially expressed genes have only six time points (the zero time point was removed from the short gene expression profiles). If the network order is greater than four, the term $|C(x, y, w, v, \dots)|$ in the denominator of Equation 3.3 is equal to 0, which the $I(x, y|w, v, ...)$ will be undefined.

To generate an appropriate structural network for clustering, the cut-off thresholds for edge removal are set deliberately for each applied dataset. First, the MI between all possible gene-pairs are computed for a set of G, the cut-off threshold for generating the zero-order network is based on the value distributions of normalized MI values (range from 0 to 1). The purpose is to ensure that at least 80% of edges in the complete network are kept for generating of higher-order networks. Later on, a even smaller threshold is set for generating the first- to fourth-order networks based on CMI values. This can prevent too many edges being removed from the network, which might generate a too-sparse network that cannot further perform the identification of functional modules. For example, in information theory, $CMI(m, n|z)$ is always smaller than or equal to $MI(m, n)$, when conditional dependence between two genes is measured given the third gene. Therefore, the threshold value is decreasing for generating the first- to fourth-order networks based on CMI value between genes.

3.3.2 Gene clustering with ClusterONE

After the gene co-expression network is generated, the next step is to identify a stress-responsive cluster/module in this network. A previous algorithm called ClusterONE (Nepusz *et al.*, 2012) is referred for this purpose. The original ClusterONE algorithm is for identifying protein complexes in a PPI network, which is considered to be one of the best algorithms for the identification of protein complexes so far. ClusterONE method is based on the assumption that the proteins in a protein complex are more densely-connected to each other than they are to the rest of the proteins in the PPI network has proved to be true. Therefore, we

believe the concept of ClusterONE can also be used to identify gene clusters/modules in the generated gene co-expression network. This is because genes involved in the same regulatory pathway for a particular function also tend to form a densely-connected sub-network when compare them to the rest of the genes in gene co-expression network (Sun *et al.*, 2011).

The identification of the cluster/module relies on the calculation of cluster cohesiveness score, which measures the likelihood of a gene cluster. The cluster cohesiveness score, denoted as $f(V)$, can be calculated as follows:

$$f(V) = \frac{w^{\text{in}}(V)}{w^{\text{in}}(V) + w^{\text{bound}}(V) + p(|V|)}, \tag{3.5}$$

since the generated gene co-expression network is an unweighted network, $w^{\text{in}}(V)$ represents the number of edges between genes within the group V, $w^{\text{bound}}(V)$ represents the number of edges that connect genes in the group V to those in the rest of network. $p(|V|)$ is the penalty term that represents any uncertain gene boundary connections in the V, which the $p(|V|)$ should be greater than 0. Simply put, $p(|V|)$ assumes that each gene in V has a certain number of undiscovered boundary connections associated with it; therefore, $p(|V|)$ needs to be considered for the correction of cluster cohesiveness score. As the majority of the genes and interactions have been discovered already in a well-studied organism genome, such as Arabidopsis genome, the value of $p(|V|)$ would be very close to 0.

The default ClusterONE algorithm first selects a gene with the highest number of connections in the generated gene co-expression network as a seed. The seed is grown to a cluster by a greedy procedure until the value of cluster cohesiveness score is greater than a predefined value. However, by default, cluster cohesiveness scores greater than 0.3 produce dense-clusters, which contains only 10 to 20 genes; whereas the clusters with cohesiveness scores smaller than 0.3 are not considered as dense-clusters. Therefore, a new method is developed for the seed selection for better adaptation to the applied datasets.

The modified ClusterONE method consists of two steps. First, a sub-network from the

generated gene co-expression network is obtained, then a greedy procedure is applied to the sub-network to identify a gene cluster/module. To obtain the sub-network, instead of selecting the gene, which has the highest connectivity among all of the genes in the gene co-expression network, the seed is selected from a list of annotated stress-responsive genes. This can be done by mapping the list of annotated stress-responsive genes to the inferred gene co-expression network. Then, the seed gene is selected from the list of annotated stress-responsive genes, which has the highest connectivity in the gene co-expression network. For example, a total of 241 annotated genes assigned to **GO term: Response to cold** in *A. thaliana* are listed according to *A. thaliana* GO annotation file downloaded from TAIR. In order to identify cold-responsive module from the cold stress expression dataset, these 241 cold-responsive genes are mapped to the inferred gene co-expression network. The proposed method selects a seed from the list of cold-responsive genes, which has the highest connectivity in the gene co-expression network, in comparison with the rest of other genes from these 241 cold-responsive genes.

The next step is to grow sub-network from the selected seed gene. The basic procedure is to successively identify neighbouring genes of the seed gene. The method first includes genes that are directly connecting to the seed in the generated gene co-expression network. It is assumed that these included directly stress-responsive genes, as they directly connect to the highly connective seed in the gene co-expression network. Then, another set of neighbouring genes, which directly connect to the directly stress-responsive genes previously identified in the gene co-expression network, are included. These additional genes are assumed to be indirectly stress-induced genes, which are regulated by directly stress-responsive genes. This sub-network growing procedure is named as "second-order neighbouring search" method.

The demonstration of how the sub-network was grown is illustrated in Figure 3.2. In Figure 3.2, the red node represents the seed gene with the highest connectivity among a list of stress-responsive genes, when map to the gene co-expression network. Blue nodes are neighbouring nodes of the seed gene that have high probabilities of being directly stress-responsive genes. Brown nodes are neighbouring nodes of directly stress-responsive genes, which are supposed

to represent indirectly stress-responsive genes induced by directly stress-responsive genes (blue nodes).

After the sub-network was obtained, Equation 3.5 is used to calculate the cluster cohesiveness score for identifying a gene cluster/module from the sub-network. The cluster cohesiveness score is calculated in a greedy procedure, which can recursively generate a set of genes from the seed. The method repetitively calculates the cluster cohesiveness score for each gene inclusion growing from the seed, and for each gene-exclusion from existing the cohesive group that has been grown, for updating of the cohesive group until the cluster cohesiveness score reaches the predefined value. The steps of the greedy procedure are as follows.

(1) The algorithm selected the seed gene, which has the highest degree of connectivity from the sub-network, then the seed was grown by including more nodes through greedy procedure by letting $V_0 = \{v_0\}$ starting with step number $t = 0$;

(2) The cohesiveness of V_{t} is calculated and assigned V_{t} to $V_{\mathrm{t+1}}$, denoted as $V_{\mathrm{t+1}} = V_{\mathrm{t}}$;

(3) For every external node v that has at least one boundary edges connected to V_{t}, the cohesiveness of $V_t^{'} = V_{\mathrm{t}} \cup \{v\}$ is calculated. If $f(V_t^{'}) > f(V_{\mathrm{t+1}})$, then assign $V_t^{'}$ to $V_{\mathrm{t+1}}$, denoted as $V_{\mathrm{t+1}} = V_t^{'}$;

(4) for every internal node v that has at least one boundary edge connected to the rest of the cohesive group, calculate the cohesiveness of $V_t^{''} = V_{\mathrm{t}} \setminus \{v\}$. If $f(V_t^{''}) > f(V_{\mathrm{t+1}})$, then assign $V_t^{''}$ to $V_{\mathrm{t+1}}$, denoted as $V_{\mathrm{t+1}} = V_t^{''}$;

(5) if $V_{\mathrm{t}} \neq V_{\mathrm{t+1}}$, let $t = t + 1$, then the whole procedure begins again from (2). If $V_{\mathrm{t}} = V_{\mathrm{t+1}}$, let V_{t} be a local optimal cohesive gene group.

A number of thresholds for the cluster cohesiveness score were tested until the best stress-responsive cluster evaluated by GO enrichment analysis was obtained. Cluster cohesiveness

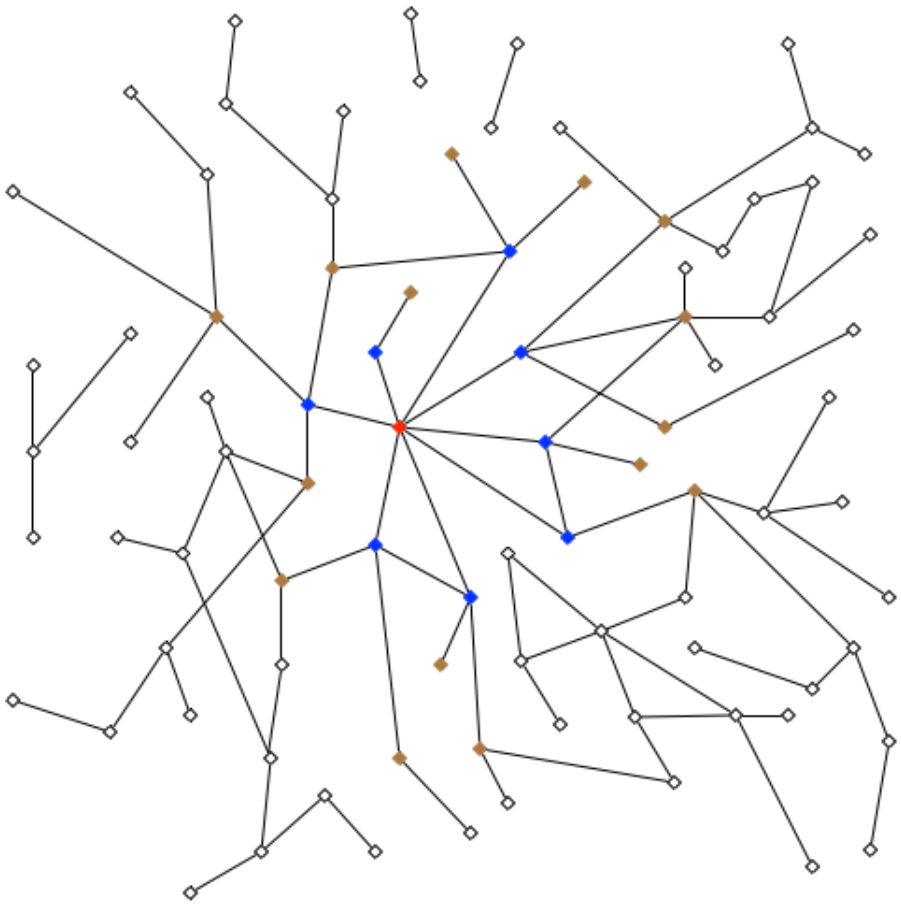

Figure 3.2: Illustration of how sub-network is generated from the seed gene.

score of 0.3 produced the best clustering results for the applied ten datasets.

3.4 Result and Discussion

3.4.1 Result of method evaluation

Gene co-expression network inference

Table 3.1 presents the cut-off thresholds for edge removal based on the value distributions of MI and CMI in the process of generating gene co-expression networks for ten datasets. The goal is to keep maximum amount of biologically contributive data while excluding data noises. Thresholds of MI are for generating the zero-order network, and thresholds of CMI are

for generating higher order networks (First- to fourth-order network). The reason that the threshold of CMI is the same for generating the first- to the fourth-order network is because the process of threshold selection is very time consuming. In order to select MI and CMI thresholds for one dataset, algorithm needs to complete one run of PCA. For each run of PCA, MI needs to be computed for all possible combination of gene-pairs, followed by computation of CMI for all possible gene-pairs for generating networks from the first- to fourth- order.

Table 3.1: Thresholds of edge deletion for MI and CMI in the process of generating gene co-expression networks for ten datasets.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
MI	0.4	0.5	0.5	0.4	0.45	0.4	0.5	0.4	0.45	0.4
CMI	0.1	0.15	0.15	0.1	0.15	0.1	0.15	0.1	0.15	0.1

Gene clustering with ClusterONE

Table 3.2: The number of annotated genes assigned to stress-related GO terms in *A. thaliana* for seed selection.

Cold GO:0009409	**Drought** GO:0009414	**Heat** GO:0009408	**Salinity** GO:0009651	**Osmotic stress** GO:0006970
254	198	96	278	256

The network-based clustering method selects a seed gene from a list of annotated stress-responsive genes, which has the highest number of connections in the generated gene co-expression network. The number of annotated genes assigned to **Cold** GO:0009409, **Drought** GO:0009414, **Heat** GO:0009408, **Salinity** GO:0009651 and **Osmotic stress** GO:0006970 in *A. thaliana* is presented in Table 3.2.

Once the seed was decided for each dataset, the "second-order neighbouring search" method was applied to obtain a sub-network for the identification of gene cluster/module using the greedy procedure for calculation of cluster cohesiveness score. The cluster cohesiveness score of 0.3 was set to exclude false positives generated by noise from inclusion in the sub-network. A penalty of 2 was set for $p(|V|)$, as the *A. thaliana* has been well-studied.

Genes remained in the network after exclusion of noise through the greedy procedure were considered to comprise the final gene cluster/module for the dataset. The final gene cluster/-module was evaluated by GO enrichment analysis, and the clustering results of 10 *A. thaliana* datasets from both my proposed method and ClusterONE were summarized in Tables 3.3, 3.4, 3.5 and 3.6.

Table 3.3: Clustering results of *A. thaliana* root tissue datasets from the network-based clustering method.

My method (Root)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	381	317	466	415	370
# of genes under GO	24	19	34	27	20
Percentage of genes under GO	6.3%	6.0%	7.3%	6.5%	5.4%
p-value of GO	2.31×10^{-14}	4.99×10^{-15}	8.39×10^{-13}	5.46×10^{-21}	4.78×10^{-17}
Percentage of indirectly stress-responsive genes	63.2%	71.5%	72.3%	70.0%	63.4%
Percentage of total stress-responsive genes	69.5%	77.5%	79.6%	76.5%	68.8%

Table 3.4: Clustering results of *A. thaliana* shoot tissue datasets from the network-based clustering method.

My method (Shoot)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	271	306	413	375	362
# of genes under GO	16	19	33	27	21
Percentage of genes under GO	5.9%	6.2%	8.0%	7.2%	5.8%
p-value of GO	4.57×10^{-15}	1.51×10^{-13}	4.72×10^{-18}	1.52×10^{-24}	4.90×10^{-13}
Percentage of indirectly stress-responsive genes	62.5%	59.2%	67.0%	71.5%	56.8%
Percentage of total stress-responsive genes	68.4%	65.4%	75.0%	78.7%	62.6%

Table 3.5: Clustering results of *A. thaliana* root tissue datasets from ClusterONE.

ClusterONE (Root)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	24	32	15	23	29
# of genes under GO	1	1	0	0	0
Percentage of genes under GO	4.2%	3.1%	0%	0%	0%
p-value of GO	1.24×10^{-1}	1.03×10^{-1}	1	1	1
Percentage of indirectly stress-responsive genes	8.3%	3.1%	0%	0%	0%
Percentage of total stress-responsive genes	12.5%	6.2%	0%	0%	0%

Table 3.6: Clustering results of *A. thaliana* shoot tissue datasets from ClusterONE.

ClusterONE (Shoot)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	15	26	19	17	26
# of genes under GO	0	1	0	0	2
Percentage of genes under GO	0%	3.8%	0%	0%	7.7%
p-value of GO	1	6.5×10^{-1}	1	1	1.91×10^{-1}
Percentage of indirectly stress-responsive genes	0%	3.8%	0%	0%	0%
Percentage of total stress-responsive genes	0%	7.6%	0%	0%	7.7%

For dataset generated from cold-treated *A. thaliana*, **GO:0009409**, response to cold, was the target GO term used for testing the result effectiveness of the proposed method in stress-responsive cluster and similarly the datasets from other stress-treated samples: **GO:0009414**, response to water deprivation for data from drought-stressed samples; **GO:0009408**, response to heat for data from heat-stressed samples; **GO:0009651**, response to salt stress for data from salt-stressed samples and **GO:0006970**, response to osmotic stress for data from osmotic-stressed samples. The network-based clustering method was compared with previous ClusterONE algorithm in terms of GO *p-values*, the percentages of directly stress-responsive genes, the percentage of indirectly stress-induced genes and total stress-related genes in the obtained cluster. As Tables 3.3, 3.4, 3.5 and 3.6 show, the target GO *p-value* for the same dataset is much smaller for the network-based clustering method than the original ClusterONE method. All the GO *p-values* from the network-based clustering method are smaller than 10^{-6}, which are considered to be significant clustering results (Erik *et al.*, 2013). The original ClusterONE generated non-sense clusters, which contain only 10 to 20 genes. It is obviously to conclude that my proposed method is able to more effectively categorize same functional genes into clusters.

3.4.2 Cross-stress comparison

In addition to the evaluation of clustered genes for each dataset, the cross-stress comparison was performed to identify commonly expressed stress-responsive genes among all five abiotic stress conditions, so called cross-talk genes. The cross-stress analysis reveals the cross-talk of genes to diverse abiotic stresses (cold, drought, heat, salt and osmotic stress) in both root and shoot tissues at the seedling stage of *A. thaliana*. The obtained list of cross-talk genes has high potential to be abiotic stress tolerant genes, when any types of abiotic stresses are presented. These cross-talk genes are presented with their gene identifiers and their molecular functions or biological processes in Tables below. References confirm that the identified cross-talk genes associated functions are indeed involved in the regulation for abiotic stress defense in *A. thaliana*.

Table 3.7: A list of cross-talk genes under five abiotic stresses from *A. thaliana* root tissue at the seedling stage by network-based approach.

At1g05100	**member of MEKK subfamily:** ATP bind, kinase activity, protein kinase activity, protein phosphorylation, transferase activity.(Yin *et al.*, 2013; Abwao, 2012)
At1g02930	**Glutathione transferase:** response to oxidative stress, response to salt stress, response to water deprivation (locates in root, seedling development stage). (Sappl *et al.*, 2009)
At4g25380	**stress-associated protein 10:** cellular response to cold, response to heat, response to salt stress, response to high light intensity, response to hydrogen peroxide, response to manganese ion, response to meta ion, response ti nickel cation, response to zinc ion (Sappl *et al.*, 2009).
At2g25080	**Glutathione peroxidase: GPX1:** glutathione peroxidase activity, response to oxidative stress. (Glombitza *et al.*, 2004; Sugimoto and Sakamoto,1997; Rodriguez Malia *et al.*, 2003)
At1g08920	**ESL1:** carbohydrate transmembrane transporter activity, response to abscisic acid stimulus, response to salt stress, response to water deprivation.

At4g15300	**cytocchrome P450 family gene: oxidoreductase activity.** (Glombitza *et al.*, 2004)
At3g23240	**ATERF1(ERF1):** defense response, ethylene mediated signalling pathways, jasmonic acid mediated signalling pathway, sequence-specific DNA binding transcription factor activity. (Cheng *et al.*, 2013)
At1g01140	**CBL:** kinase activity, response to cold, response to mannitol stimulus, response to salt stress. response to wounding. (Yin *et al.*, 2013)
At1g29395	**Cold regulated 314 inner membrane 1:** response to cold, response to water deprivation, response to hyperosmotic salinity, response to salt stress.
At1g02500	**ATSAM1:** response to iron ion, response to hyperosmotic, response to salt stress, response to temperature stimulus, response to wounding (locates in root, at seedling development stage)
At1g02450	**NIMIN1:** MAPK cascade, regulation of defense response, regulation of plant-type hypersensitive response, regulation of protein dephosphorylation, regulation of transcription. (Yin *et al.*, 2013)
At5g50360	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At5g24600	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)

Table 3.8: A list of cross-talk genes under five abiotic stresses from *A. thaliana* shoot tissue at the seedling stage by network-based approach.

At1g29395	**Cold regulated 314 inner membrane 1:** response to cold, response to water deprivation, response to hyperosmotic salinity, response to salt stress.
At1g47128	**RD21:** cysteine-type peptidase activity, response to hyperosmotic, response to salt stress, response to temperature stimulus, response to water deprivation.
At2g25080	**Glutathione peroxidase: GPX1:** glutathione peroxidase activity, response to oxidative stress. (Glombitza *et al.*, 2004; Sugimoto and Sakamoto,1997; Rodriguez Malia *et al.*, 2003)

At4g15300	**cytocchrome P450 family gene:** oxidoreductase activity. (Glombitza *et al.*, 2004)
At2g47180	**GolS1:** carbohydrate biosynthetic process, galactosyltransferase activity, response to abscisic acid stimulus, response to cold, response to heat, response to high light intensity, response to oxidative stress, response to salt, response to water deprivation, transferase activity.(Abwao, 2012)
At1g01620	**PIP1c:** cellular response to iron ion starvation, response to salt stress, response to water deprivation (at seedling development stage).
At1g02730	**ATCSLD5:** glucosyltransferase activity, regulation of cell proliferation, response to osmotic stress, response to salt stress, response to water deprivation (locate in the shoot system development).(Abwao, 2012)

Conclusively, results shown in the above tables illustrate that our proposed network-based clustering method can not only effectively perform cluster analysis on short gene expression profiles, but can also produce results with significant biological meaning involves in general abiotic stress defense, known as cross-talk genes. These identified cross-talk genes were subjected to further validation to confirm their involvement in the key regulation of abiotic stress. Additionally, it was concluded that the gene expression of abiotic stressed samples from the *A. thaliana* root tissue appeared to be more homogenous than the gene expression of abiotic stressed samples from the *A. thaliana* shoot tissue, because the stressed datasets generated from root tissue samples identified more cross-talk genes under five abiotic stressed conditions than datasets generated from shoot tissue. Among the identified abiotic-stress tolerant genes in the root tissue, two candidate genes are novel genes without annotation information. There are mutant lines available for these two novel genes according to the search of gene mutants in TAIR. Plant seeds can be ordered for those mutant lines in order to study the significance of identified novel genes for plant growth under abiotic stress conditions. Therefore, the list of identified cross-talk genes among all five abiotic stress conditions have provided very useful information for biologists to study gene functions and to identify genes associated with key regulation in abiotic stress response.

3.5 Conclusion

The network-based clustering method was presented in this chapter. This method first generated a gene co-expression network showing interdependent relationships between genes in short gene expression profiles based on CMI. Next, a cluster cohesiveness score was calculated in a greedy procedure to effectively identify a cluster of genes responds to stress defense from the extracted sub-network. Both the network-based clustering method and the original ClusterONE algorithm were employed on 10 datasets for comparison of methods effectiveness. Results showed that the network-based clustering method produced significant results in terms of 1) GO *p-values* for directly stress-responsive genes, and 2) percentages of total stress-related genes in the obtained stress-responsive clusters for ten datasets. Further more, cross-stress comparison identified cross-talk genes under five abiotic stress conditions from both root and shoot tissues of *A. thaliana*. These identified candidate genes were subjected to further validation to confirm their involvement in the key regulation of abiotic stress, which will help further analyses of fundamental abiotic stress regulation that lead to a better discovery of abiotic stress defense system.

Chapter 4

PPI-integrated clustering method

4.1 Overview

A genome encodes numerous proteins with multiple biological functions. As whole genome are being sequenced, and their functions are analyzed, it is necessary to develop clustering methods to classify genes into multiple clusters such that each represents a biological function. This chapter proposes a clustering method for more specifically clustering genes based on their short gene expression profiles with integrating PPI data.

In order to better cluster gene expression profiles, a number of methods have been developed to deal with time-series gene expression data in different lengths. A tool called STEM (Short Time-series Expression Miner) (Ernst and Joseph, 2006) is designed especially for clustering short time-series gene expression profiles without requiring prior knowledge. It predefines various model profile patterns, each of which represents a cluster mean profile that is independently from the experimental gene expression data. Construction of model profile patterns is based on the number of time points in gene expression profile and the unit change between any two consecutive time points. Each gene in a dataset is assigned to a matching model profile pattern according to the similarity of a gene expression profile to the model profile pattern of a cluster (Ernst and Joseph, 2006). It is clever to predefine model profile patterns for overcoming the short time-series problem, as each pair of consecutive time points is taken into consideration for building up model profiles, which increases possible numbers of pattern combinations.

In order to further improve clustering effectiveness, other types of data than gene expression

profiles can be integrated into the cluster analysis. To incorporate other types of data in cluster analysis, it is critical to select appropriate data to maximize data contribution. Since the purpose of gene cluster analysis is to effectively group functionally related genes into clusters, it would be helpful if protein information can be integrated, as genes only carry out functions when they are expressed at the protein level. However, it is not appropriate to integrate amino acid sequence alignment information for the estimation of pairwise similarity between gene-pairs (White *et al.*, 2010), as protein is only functional in complex structural level, whereas amino acid sequence is in primary sequential level. Combination of such inappropriate heterogeneous data might produce a poor clustering results. A Bayesian model based algorithm has been developed in the paper (Kirk *et al.*, 2012), which is able to generate clustering result by integrating heterogeneous data, but the effectiveness of this method has no significant improvement.

As it is known that proteins rarely act alone, but interact to form protein complexes that are capable to carry out functional performances. Based on assumption that genes coding interacted proteins should be in the same functional cluster, the PPI data should be helpful in discovering biological patterns from short time-series gene expression profiles. A PPI network can be generated from PPI data, which the network represents physical interactions among proteins, with nodes representing proteins and edges representing interaction relationships between proteins. For cluster analysis, if there is an edge connecting two proteins in a PPI network, there is a high possibility that two proteins have interacted with each other sometimes during a biological process. Therefore, the protein connectivity information in PPI data can assist to group similar functional genes into the same cluster in cluster analysis. The paper (Sun *et al.*, 2011) has used three typical clustering algorithms to decompose PPI network into dense sub-networks as clusters, which has achieved contributive results by using PPI data alone.

In this thesis, a novel method to cluster short time-series expression profiles with integrating PPI data was proposed. The schematic of this proposed method is illustrated in Figure 4.1 within the dashed line box, and the brief steps of the method are summarized as follows:

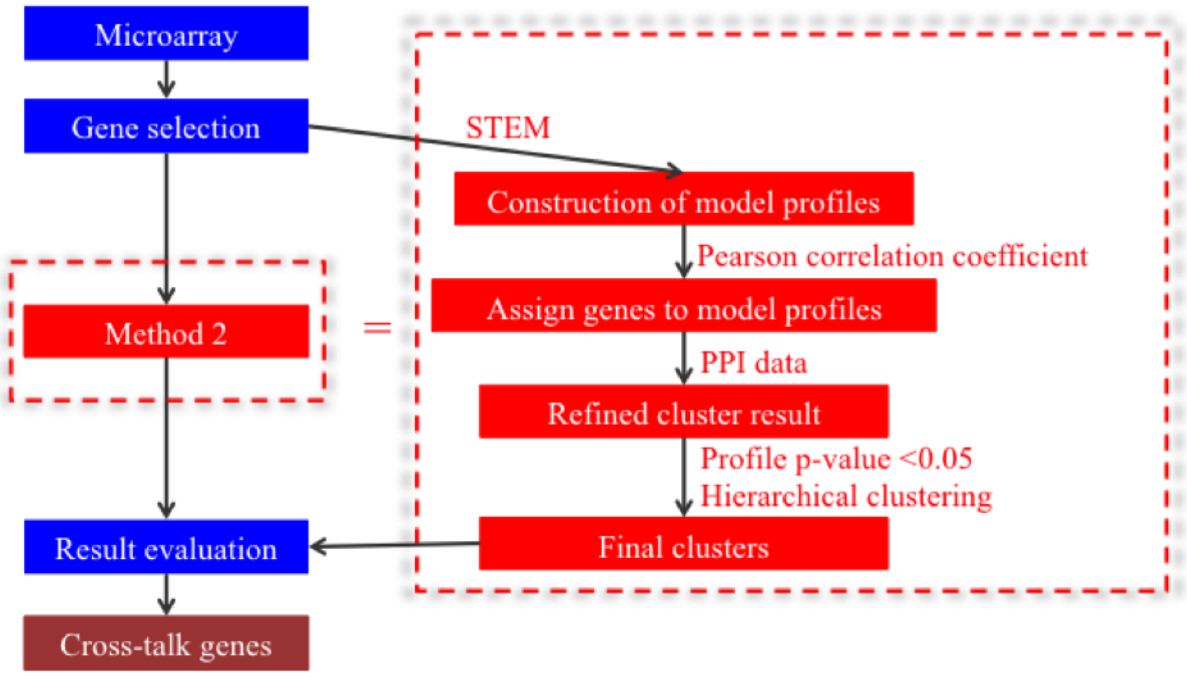

Figure 4.1: Schematic of the PPI-integrated clustering method.

(1) Generate a number of predefined profile patterns according to the number of data points in the gene expression profiles and assign each gene to one of predefined profile patterns which its expression profile pattern is the most similar;

(2) Integrate PPI data to refine the initial clustering result from step one;

(3) Combine similar clusters obtained from step two to gradually reduce the number of clusters by hierarchical clustering method.

4.2 Method

4.2.1 Model profile pattern construction

The logic of setting up model profile patterns prior to cluster analysis was adopted from previous short time-series gene clustering algorithm, called STEM (Ernst *et al.*, 2005). The

method constructed a fixed number of model profile patterns with various expression patterns. A gene was assigned to one of constructed model profile patterns to which its expression profile was the most similar. The construction of model profile patterns was decided by two variables, which were the number of time points for the gene expression profile and the number of unit changes between two consecutive time points. The number of model profile patterns was calculated as follows (Ernst *et al.*, 2005):

$$P = (2U + 1)^{T-1} - 1, \tag{4.1}$$

where P is the number of all possible model profile patterns, T is the number of time points according to the gene profile in the dataset, and U is the number of unit changes between two consecutive time points.

All model profile patterns start at 0, increase or decrease in an integral unit number that is equal to the value of U. The model profile pattern with constant values throughout all time point was excluded.

Typically, three different expression states between two consecutive time points were considered, which were increasing, decreasing or constant. We set $U = 1$ as the unit change between every consecutive time points, and $T = 7$ as the number of time points according to the gene profiles of *A. thaliana* dataset. Therefore, the vector of a model profile, which represents the down-regulated gene pattern during the entire experiment duration would be $(0, -1, -2, -3, -4, -5, -6)$, which each time point in comparison with its previous time point has a unit change of -1. Therefore, PCC value between a gene expression profile and a model profile pattern was computed by measuring the similarity between the unit change pattern for a model profile pattern and expression change pattern for a gene expression profile. As a result, there were a total of 728-model profile patterns defined for each dataset under considerations. Each differentially expressed gene (> two-fold) was assigned to the most similar model profile pattern according to PCC between each gene profile and model profile patterns.

As the PCC value falls between -1 to 1, where 1 represents positive correlation, -1 represents negative correlation and 0 represents no correlation between two variables. A gene expression profile pattern was assigned to the closest model profile which their PCC value was the largest.

The PCC value between the expression change pattern of $(x_1, x_2, x_3, ...x_n)$ for a gene expression profile and the unit change pattern of $(y_1, y_2, y_3, ...y_n)$ for a model profile pattern was calculated as follows:

$$PCC(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.2}$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

For the identification of significant model profile patterns, given a set of differentially expressed genes G and a set M of total constructed model profiles; each gene $g \in G$ was assigned to $m_i \in M$ if $PCC(g, m_i)$ was the maximum over all $m \in M$. For model profile m, the number of genes assigned to it was denoted as T_m.

Next step is to identify model profile patterns that were significantly enriched in G. To do this, the null hypothesis was that the number of genes assigned to a model profile according to the PCC was equal to the expected number of genes assigned to the model profile pattern randomly. Permutation was used to estimate the expected number of genes assigned to a model profile pattern at random. In my method, permutation had shuffled time points of G 50 times, in which each possible permutation could generate a set of random genes G_n $(n = 1, \ldots, 50)$ that could be assigned to their closest model profile pattern. Let the number of genes assigned to model profile pattern m $(m \in M)$ in the nth permutation be s_m^n. We set $s_m = \sum_{i=50} s_m^n$. Then $E_m = s_m/50$ was the estimation of the expected number of genes assigned to model profile pattern m.

Since each $g \in G$ was assigned to one of M model profile patterns, is was assumed that the number of genes assigned to each model profile pattern was distributed as binomial random variable as $X \sim Bin(|G|, \frac{E_m}{|G|})$. Therefore, the *p-value* of seeing T_m genes assigned to model

profile m was $P(X \geq T_m)$. Model profile patterns with a *p-value* smaller than $\alpha = 0.05$ were considered to be significant model profile patterns.

As the significance levels of M model profiles were tested, the *p-value* needed to be corrected for multiple comparisons. Therefore, Bonferroni correction was applied to correct each model profile *p-value*, by dividing the value of α by M, thus model profile patterns with $P(X \geq T_m) < \alpha/M$ were considered to be significant model profile patterns (Ernst *et al.*, 2005).

4.2.2 PPI incorporation to refine patterns

If the total numbers of model profile patterns were denoted as k, clusters could be represented by $P_1, P_2, P_3, \ldots, P_k$, with various numbers of genes assigned to them. In step two, genes assigned to clusters were refined by integrating PPI data downloaded from the TAIR database. First of all, gene IDs assigned to profiles were transformed into corresponding protein IDs to look up matching edges in PPI data for cluster refinement. The detailed refinement steps are described as Algorithm below.

Input: A set of differentially expressed genes fit into model profiles P_1, P_2, P_3, P_k (the target gene in the model profile is denoted as g_{ij}.

A PPI network G, in which each node is represented by proteins encoded by the gene.

Output: A modified assignment of gene profiles P_1', P_2', P_3', P_k'.

for i = 1 to k (where k is the number of profiles)

for j = 1 to $|P_\mathrm{i}|$(where $|P_\mathrm{i}|$ is the gene size of profile i)

let $\mathrm{N}_1 = \mathrm{g}_{\mathrm{ij}}$ (in which node 1 is represented by g_{ij})

Number of matching edges in each profile for target g_{ij} is initialized to 0, denoted as (S_1, S_2, S_3, $S_k \leftarrow 0$)

for p = 1 to k

for q = 1 to $|P_i|$

let $N_2 = P_{pq}$ (in which node 2 is represented by g_{pq})

if (N_1, N_2) ϵ E(G) (if any N_1, N_2 edge matches with an edge in the PPI network)

$S_p = S_p+1$ (add 1 score to the p^{th} profile)

end

end

end

(Pmax Imax)= max (S_1, S_2, S_3, S_k) (where P_{max} is the maximal profile score, I_{max} is the profile index of p_{max})

stores I_{max} for g_{ij} in this run in cell, continue with the next loop

end

end

At the end, move all the target g_{ij} to their corresponding $I_{max}{}^{th}$ profile.

After the memberships of genes in clusters had been refined by PPI data, *p-values* of 728 model profiles patterns were calculated again according to binomial distribution of the expected number of genes assigned to model profile patterns for selection of significant profile patterns. Genes in profile patterns with *p-value* less than 0.05 were exported as a cluster of input genes in step 3.

4.2.3 Hierarchical clustering for final patterns

To further improve the significance of discovered profile patterns, A hierarchical clustering method was applied to group sets of significant profile patterns into larger clusters. Hierarchical clustering produced tree like structure showing similarity relationships among clusters (Szekely and Rizzo, 2005). It could be used for not only clustering gene expression profiles, but also for clustering any datasets in which similarity relationships could be defined. In contrast to k-mean clustering, it avoided defining cluster numbers priori to cluster analysis. Instead, the number of obtained clusters in hierarchical clustering depended on the cut-off threshold (Yeung and Ruzzo, 2001), or through visualizing the dendrogram structure to draw cluster boundaries manually. The bottom-up hierarchical clustering algorithm started with N sample profile patterns. Any two most similar profile patterns were grouped to form a new cluster. In subsequent steps, similarities of the newly formed cluster and remaining profile patterns were calculated to iteratively merge profile patterns; until no more profile patterns could be merged, and a dendrogram was formed. There are complete-linkage clustering (Defays, 1977), average-linkage clustering (Sokal and Michener, 1958) and single-linkage clustering (Sibson, 1973) for merging profile patterns using their generated PCC values.

In my proposed method, significant profile patterns with *p-values* less than 0.05 after PPI refinement from step 2 were used as input profile patterns for hierarchical clustering. The similarity between genes in one profile pattern to a gene in another profile pattern was defined in terms of PCC value. The similarity between two profile patterns was calculated by taking the PCC average of all genes in one profile pattern to all genes in another profile pattern, known as the average-linkage method. The PCC value between any two profile patterns was used to subsequently merge closely related profile patterns by hierarchical cluster analysis to form dendrogram structure representing profile pattern relationships. Cluster boundaries were drawn by visualizing dendrogram structure, which clearly separated one cluster from others. After clusters are decided, GO has been applied to validate results from my proposed method.

4.3 Result and Discussion

Generally, plants are growing under various weathers and geological conditions. It is necessary for plants to have stress adaptive responses for preventing them from harsh conditions. At the molecular level, adaptation of plants to various stresses depends upon activation of cascades of cellular pathways in signal transduction, as well as expression of stress-responsive genes (Huang *et al.*, 2012).

To demonstrate the effectiveness of my proposed method, It was applied to 10 publicly available *A. thaliana* datasets.

4.3.1 Selection of significant profile patterns

Table 4.1: Numbers of significant model profile patterns after PPI refinement out of total constructed model profile patterns for 10 *A. thaliana* datasets.

n / N	Root	Shoot
Cold	121/728	152/720
Drought	26/711	52/717
Heat	48/720	43/721
Salinity	134/727	101/727
Osmotic stress	149/727	182/727

Table 4.1 shows the number of significant model profile patterns (p-value <0.05) selected after PPI refinement out of total constructed model profile patterns for each applied *A. thaliana* short microarray gene expression dataset, where: n/N = numbers of significant profile patterns / numbers of total constructed model profile patterns.

There should be 728-model profile patterns generating from all ten datasets according to Equation (1), as $U = 1$ and $T = 7$. However, I obtained different numbers of initial clusters from each dataset. The missing initial clusters were due to the fact that there were no genes

assigned to their corresponding model profile patterns. Therefore, the empty initial clusters were discarded.

4.3.2 Clustering results of *A. thaliana* datasets

The proposed novel method had applied to 10 *A. thaliana* datasets, five of which were taken from root tissues treated under 5 different abiotic stresses, and another five of which were taken from shoot tissues also treated under the same 5 different abiotic stresses as described in Chapter 2. The purpose is to exam the method effectiveness for clustering short gene expression profiles into biological functional clusters. Figure 4.2 presents the obtained clusters, as well as their corresponding functions for each of ten datasets. The cluster marked in red corresponds to the target stress-responsive cluster obtained for each dataset. The evaluation results for method effectiveness, which will be presented in later of this section, are mainly focusing on the analysis of these stress-responsive clusters in red.

	Cold	Drought	Heat	Salt	Osmotic
Root	1. Cellular metabolic process 2. Program cell death 3. N/A 4. Response to cold 5. Response to stimulus 6. Response to stress 7. N/A 8. Sulfur compound biosynthesis process	1. Response to water deprivation 2. Response to chitin 3. N/A 4. Root cell differentiation 5. Program cell death	1. Metabolic process 2 Microtubule-based process 3. Sugar metabolic process 4. Response to heat	1. Chromatin organization 2. Cellular process 3. Program cell death 4. Response to salt 5. Response to temperature	1. Methylation 2. Epidermal cell differentiation 3. Response to osmotic 4. Pollen tube growth 5. Response to light intensity
Shoot	1. Cellular metabolic process 2. N/A 3. Program cell death 4. Sulfur compound biosynthesis process 5. N/A 6. N/A 7. Response to cold 8. Cell differentiation	1. Program cell death 2. Response to water deprivation 3. Response to chitin 4. N/A 5. Cell differentiation	1. Sugar metabolic process 2. N/A 3. Response to heat 4. Metabolic process 5. Program cell death	1. Response to temperature 2. Response to salt 3. Cell differentiation 4. Program cell death 5. Chromatin organization 6. Metabolic process	1. Response to osmotic 2. Response to stimulus 3. Epidermal cell differentiation 4. Program cell death 5. N/A 6. Methylation

Figure 4.2: Clustering results for ten *A. thaliana* datasets.

4.3.3 Performance improvements through my proposed clustering method

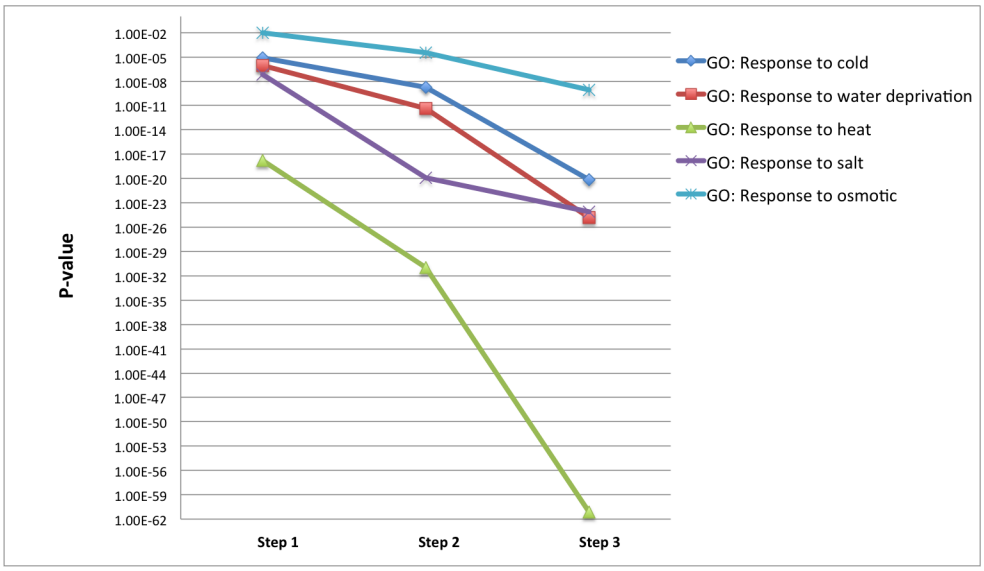

Figure 4.3: Reduction of GO *p-values* from step 1 to step 3 of the PPI-integrated clustering method for five datasets generated from the *A. thaliana* root tissue.

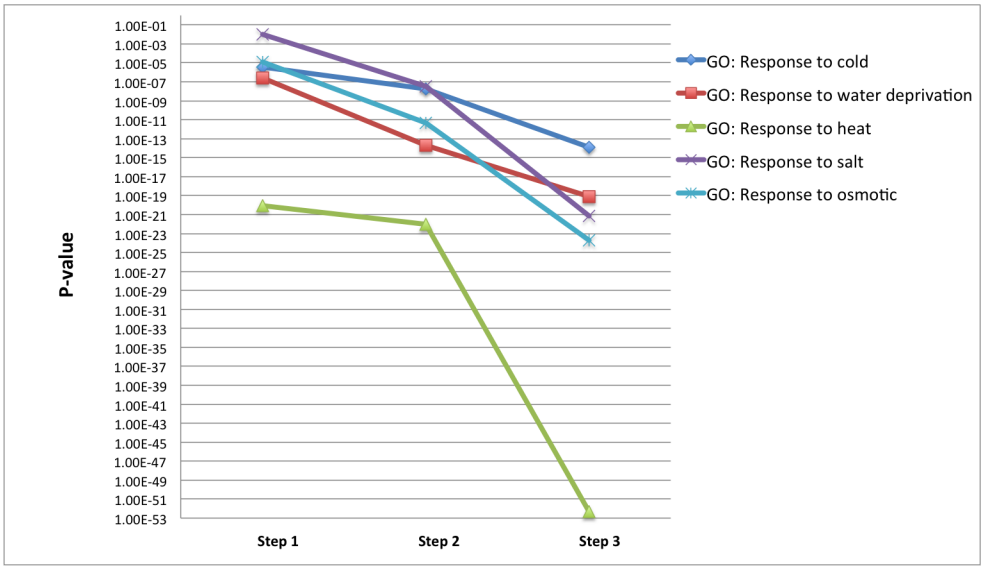

Figure 4.4: Reduction of GO *p-values* from step 1 to step 3 of the PPI-integrated clustering method for five datasets generated from the *A. thaliana* shoot tissue.

Figures 4.3 and 4.4 illustrate the reduction of *p-values* for GO: 0009409 *p-value*, or response to cold (dark blue); GO: 0009414, or response to water deprivation (red); GO: 0009408, or response to heat (green); GO: 0009651, or response to salt (purple) and GO: 0006970, or response to osmotic stress (light blue). In step 1, genes with expression change greater than two-fold in at least one time point are inputted for calculating the significance level of target GO term. In step 2, genes in select significant profiles are used for input to calculate the target GO *p-value*. In step 3, genes in the obtained stress-responsive cluster are used for input to calculate the target GO *p-value*. It is shown that the proposed clustering method can effectively reduce the target GO *p-value* step after step for all ten datasets.

Figure 4.3 plots a gradual decreased target GO term *p-values* for five *A. thaliana* date sets deleted to the root, while Figure 4.4 presents those for for five *A. thaliana* date sets related to the shoot. The plots of these two figures have approved the significance to carry out each step of the proposed novel clustering method, as it is shown that all datasets had their GO *p-values* gradually decreased throughout steps. As *p-value* is associated with test of significance level, wherein the lower the *p-value*, the more significant a cluster of stress-related genes enriches the target GO term. The decreased GO *p-values* throughout steps clearly reflected the increased method effectiveness for categorizing functional genes into biologically significant clusters.

4.3.4 Comparison with STEM

In a complex cell environment, most genes are coordinated and cooperate with one another for functional performance. In most cases, over-expression of one gene can either stimulate or repress the expression of another gene, in either an acyclic or a loop relational form. Therefore, it is difficult to conclude exactly which genes are involved in the stress defense unless the entire pathway is considered. Theoretically, clustering based on gene expression profiles can effectively group genes into same functional cluster, since coordinated genes tend to over-express or under-expressed at the same pattern.

Table 4.2: Clustering results of *A. thaliana* root tissue datasets from the PPI-integrated clustering method.

My method (Root)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	734	445	1477	1663	464
# of genes under GO	66	48	109	125	44
Percentage of genes under GO	9.0%	10.8%	7.4%	7.5%	9.5%
p-value of GO	1.20×10^{-23}	4.53×10^{-18}	6.89×10^{-62}	7.71×10^{-25}	8.73×10^{-10}
Percentage of indirectly stress-responsive genes	77.7%	71.7%	84.1%	73.0%	61.7%
Percentage of total stress-responsive genes	86.7%	82.5%	91.5%	80.5%	71.2%

Table 4.3: Clustering results of *A. thaliana* shoot tissue datasets by the PPI-integrated clustering method.

My method (Shoot)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	1100	543	882	565	1676
# of genes under GO	72	46	82	66	131
Percentage of genes under GO	6.5%	8.5%	9.3%	11.7%	7.8%
p-value of GO	1.19×10^{-14}	8.13×10^{-20}	4.66×10^{-53}	7.30×10^{-22}	1.81×10^{-24}
Percentage of indirectly stress-responsive genes	74.0%	64.0%	85.3%	65.6%	72.6%
Percentage of total stress-responsive genes	80.5%	72.5%	94.6%	77.3%	80.4%

Table 4.4: Clustering results of *A. thaliana* root tissue datasets by STEM.

STEM (Root)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	810	326	508	907	532
# of genes under GO	79	31	51	92	42
Percentage of genes under GO	9.8%	9.5%	10.0%	10.1%	7.9%
p-value of GO	1.36×10^{-20}	2.60×10^{-14}	8.90×10^{-33}	8.16×10^{-25}	9.78×10^{-7}
Percentage of indirectly stress-responsive genes	62.5%	59.9%	70.2%	68.3%	61.1%
Percentage of total stress-responsive genes	72.3%	69.4%	80.2%	78.4%	69.0%

Table 4.5: Clustering results of *A. thaliana* shoot tissue datasets by STEM.

STEM (Shoot)	**Cold**	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
# of genes in cluster	823	406	519	516	1212
# of genes under GO	53	39	56	63	100
Percentage of genes under GO	6.4%	9.6%	10.8%	12.2%	8.3%
p-value of GO	5.70×10^{-10}	1.70×10^{-18}	2.46×10^{-38}	8.56×10^{-22}	1.17×10^{-19}
Percentage of indirectly stress-responsive genes	64.6%	58.8%	71.2%	62.8%	64.0%
Percentage of total stress-responsive genes	71.0%	68.4%	82.0%	75.0%	72.3%

Here, the method effectiveness is evaluated based on the identification of directly stress-responsive genes in terms of GO *p-values* and percentages in cluster, as well as the identification of indirectly stress-responsive genes in terms of their percentages in cluster, in comparison

with STEM. Tables 4.2 and 4.3 summarize the clustering results of 10 *A. thaliana* datasets from my proposed method, and tables 4.4 and 4.5 summarize the clustering results of 10 *A. thaliana* datasets from STEM.

Directly stress-responsive genes

For dataset generated from cold-treated *A. thaliana*, **GO:0009409**, response to cold, was the target GO term used for testing the result effectiveness of the proposed clustering method in stress-responsive cluster and similarly the datasets from other stress-treated samples: **GO:0009414**, response to water deprivation for data from drought-stressed samples; **GO:0009408**, response to heat for data from heat-stressed samples; **GO:0009651**, response to salt for data from salt-stressed samples and **GO:0006970**, response to osmotic stress for data from osmotic-stressed samples. My method was compared with previous STEM method in terms of GO *p-values* of directly stress-responsive genes in the obtained clusters. As Tables 4.2, 4.3, 4.4 and 4.5 show, the target GO *p-value* for the same dataset is always smaller for my proposed clustering method than for the STEM method, indicating that my method is able to more effectively categorize same functional genes into clusters.

Since the low target GO *p-value* only convinces the fact that stress-responsive cluster can significantly enrich the target GO term, The rest of genes in the stress-responsive cluster are also needed to be analyzed for biological significance. The rest of genes in each stress-responsive cluster have significant biological meaning with functions induced by genes assigned to the target GO term, which participate in indirect regulation for stress defense, so called indirectly stress-induced genes. Therefore, indirectly stress-induced genes were later being analyzed, in order to conclusively suggest the advantages of my proposed novel clustering method. This brings us to the evaluation of the indirectly stress-induced genes in the next section.

Indirectly stress-induced genes

The two diagrams in Figure 4.5 are examples of functional and regulatory genes induced by cold and drought stress, respectively; which are functions associated with stress responses as

well. Therefore, the genes in the cluster with products associated with the indicated functions should also be included in the obtained stress-responsive (**cold, drought, heat, salt or osmotic stress**) cluster.

For better adaptation and higher survival rate in harsh conditions, plants exhibit a variety of responses to abiotic stresses. It has been confirmed that the stress defense system of *A. thaliana* constitute a network that is interconnected at many signaling pathway levels (Knight and Knight, 2001). Sometimes, a portion of one signal transduction pathway can be triggered by a variety of stress conditions, due to the cross-talk gene elements in the pathway that connect multiple pathways. Therefore, there must be portions of elements that are commonly found to be over-expressed among multiple stress conditions.

For the applied 10 datasets of *A. thaliana* whose samples were treated with 5 different types of abiotic stresses (cold, drought, heat, salt stress and osmotic stress). Different abiotic stresses can activate a common biological pathway for defense response under certain conditions. For example, drought, cold and salt stressed plants can all stimulate the dehydration protection mechanism of plants. These stresses for example can activate catalase and peroxidases activities, which can protect against oxidative damage to the cell from the stresses (Colcombet and Hirt, 2008; Hrmova and Lopato, 2013; Knight and Knight, 2001; Shinozaki and Yamaguchi-Shinozaki, 1996). In metabolic profiling analysis, previous study had found that the majority of heat-shock responses were shared with cold-shock responses in *A. thaliana* (Kaplan *et al.*, 2004). As well for stresses of **drought and cold**, a previous survey of about 1,300 *A. thaliana* genes had found that the majority of cold- and drought-stress-regulated genes share stress responses (Seki *et al.*, 2001). Therefore, the above observations support the hypothesis that a common signal transduction pathway can be triggered by multiple stress conditions.

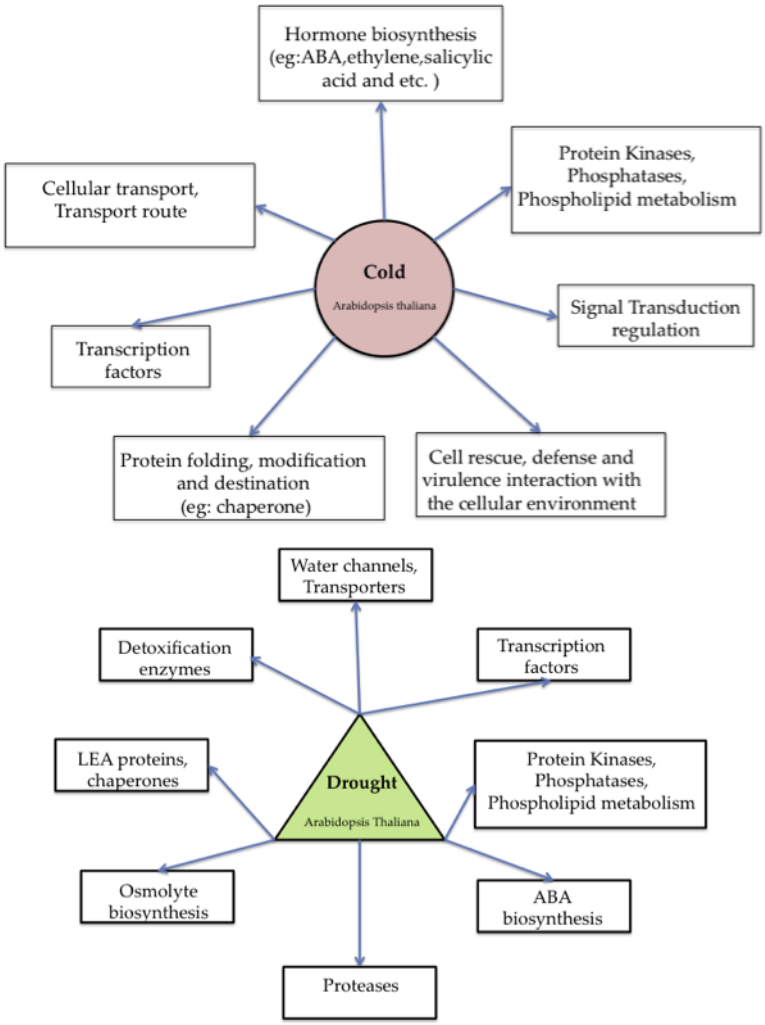

Figure 4.5: Example of cold- and drought-induced genes functions in *A. thaliana*.

Stresses in *A. thaliana* can induce both functional proteins and regulatory proteins. Such induction may lead to programmed cell death, cold tolerance, etc. (Lee *et al.*, 2005; Shinozaki and Yamaguchi-Shinozaki, 2007). Experiments for plant functional genomics have shown that multiple abiotic stresses activate transporter activity, transcription factor activity, transferase and hydrolase, all of which participate in the formation of surface structures of cells for better defense and protection from harsh conditions (Hrmova and Lopato, 2013). In addition, abiotic stresses such as cold, salt, drought, etc. are known to activate signal transductions in protein phosphorylation and protein kinase activity (Bush, 1995; Trewavas, 1999; Knight and Knight, 2001). Therefore, majority of the cold- and drought-induced gene functions as shown in Figure 4.5 can also apply to heat, salt and osmotic stresses, for analysis

of indirectly stress-induced genes in obtained stress-responsive clusters.

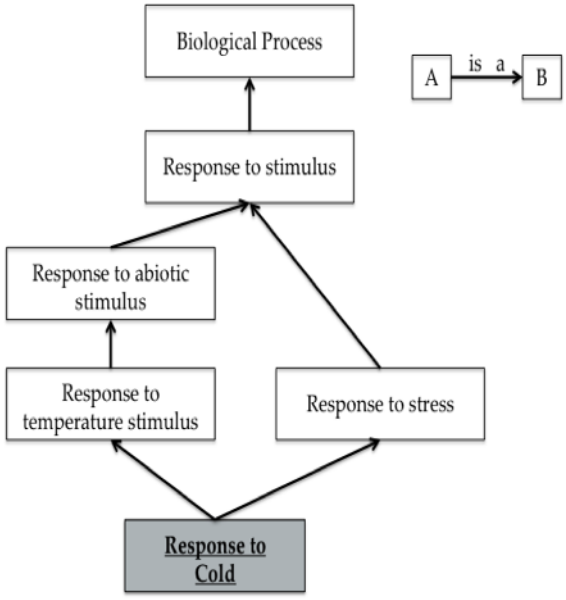

Figure 4.6: A portion of directed acyclic graph of GO term domain relations. Term box (child term) at the root of arrow belongs to term box (parent term) at the tip of arrow.

Figure 4.6 illustrates a directed acyclic graph of GO terms with more general descriptive terms located at the top of the graph and more specific descriptive terms placed at the bottom. Genes assigned to the term at the bottom of the term structure should also belong to terms above it. Therefore, it is possible to assign any indirectly stress-induced genes in the cluster to the higher GO term domains in the directed acyclic graph.

Tables 4.6 and 4.7 compare the indirectly stress-induced genes in the stress-responsive clusters obtained from both my method and STEM. Percentages of genes which are indirectly regulated by genes under the target GO term in obtained stress-responsive clusters are presented. Previous genetics and biological studies have shown that molecular functions or biological processes with descriptive terms such as kinase activity, transferase activity, signal transduction, transporter activity, hydrolase activity, etc. can control or be controlled by stress-responsive genes. Such terms can describe functional or regulatory genes, which are induced by stress-responsive genes; therefore, the functional or regulatory genes also involved

in stress defense.

Other than evaluating the indirectly stress-induced genes, identification of un-annotated genes in obtained stress-responsive cluster is also another criterion to evaluate clustering effectiveness. Each of Tables 4.6 and 4.7 contains a line indicating the percentage of genes with unknown molecular functions, where genes in this category has high possibility to consist of un-annotated genes with function associated with associated stress response. It has been shown that my proposed clustering method can obtain higher percentages of un-annotated genes in the obtained cluster for each dataset than STEM.

Percentages in tables are calculated by the following equations:

$$\% = \frac{\#\ of\ annotations\ to\ terms\ in\ the\ GO\ slim\ category \times 100}{\#\ of\ total\ annotations\ to\ terms\ in\ this\ ontology}$$

GO slim in above equation is a cut-down version of the GO containing a subset of the terns in the whole GO.

Conclusively, my proposed method shows better clustering results by obtaining higher percentages of overall stress-related genes (addition of direct stress-responsive genes and indirect stress-induced genes) in obtained stress-responsive clusters than STEM for all datasets, as shown in the last lines of Tables 4.2, 4.3, 4.4 and 4.5.

Table 4.6: Comparison of stress-associated molecular functions and biological processes in obtained clusters from datasets generated from *A. thaliana* root tissue samples by my proposed method and STEM.

Cold stress(Root)	My method	STEM
Hydrolase activity	9.05%	8.52%
Transcription factor activity	4.41%	3.34%
Unknown molecular functions	10.83%	8.64%
Protein binding	5.83%	5.58%
Signal transduction	6.90%	5.88%
Drought stress(Root)	**My method**	**STEM**
Hydrolase activity	9.59%	7.28%
Transcription factor activity	3.61%	3.21%
Unknown molecular functions	10.11%	9.87%
Development process	2.56%	1.72%
Heat stress(Root)	**My method**	**STEM**
Hydrolase activity	8.78%	7.49%
Kinase activity	5.13%	3.23%
Unknown molecular functions	15.13%	14.17%
Transporter activity	6.03%	4.95%
Transferase activity	8.73%	7.30%
Salt stress(Root)	**My method**	**STEM**
Hydrolase activity	7.77%	6.13%
Unknown molecular functions	8.79%	7.03%
Response to abiotic stress	11.35%	10.56%
Kinase activity	13.07%	11.48%
Osmotic stress(Root)	**My method**	**STEM**
Response to abiotic stress	9.49%	8.08%
Transferase activity	11.74%	10.16%
Unknown molecular functions	9.83%	8.02%
Kinase activity	5.37%	3.58%
Signal transduction [55]	4.11%	2.78%

Table 4.7: Comparison of stress-associated molecular functions and biological processes in obtained clusters from datasets generated from *A. thaliana* shoot tissue samples by my proposed method and STEM.

Cold stress(Shoot)	My method	STEM
Hydrolase activity	10.87%	9.05%
Response to abiotic stress	9.79%	8.69%
Unknown molecular functions	11.35%	10.77%
Kinase activity	6.92%	5.65%
Transferase activity	14.70%	13.61%
Drought stress(Shoot)	**My method**	**STEM**
Response to abiotic stress	16.79%	15.06%
Unknown molecular functions	14.92%	12.60%
Transferase activity	13.51%	12.97%
Signal transduction	8.37%	7.78%
Heat stress(Shoot)	**My method**	**STEM**
Response to abiotic stress	12.09%	10.13%
Transferase activity	10.28%	7.32%
Unknown molecular functions	17.77%	16.41%
Kinase activity	5.56%	3.12%
Hydrolase activity	8.76%	8.09%
Salt stress(Shoot)	**My method**	**STEM**
Transferase activity	15.04%	13.21%
Transporter activity	7.70%	6.39%
Unknown molecular functions	9.93%	8.96%
Response to abiotic stress	13.68%	12.16%
Kinase activity	6.07%	5.53%
Osmotic stress(Shoot)	**My method**	**STEM**
Hydrolase activity	9.34%	8.65%
Transcription factor activity	7.76%	2.57%
Unknown molecular functions	10.64%	8.45%
Response to abiotic stress	11.43%	9.57%
Transferase activity	13.56%	10.05%

4.3.5 Cross-stress comparison

In addition to the evaluation of clustered genes for each dataset, I also performed cross-stress comparison to identify cross-talk genes. The cross-stress analysis reveals cross-talk of responsive genes to multiple abiotic stress (cold, drought, heat, salt stress and osmotic stress) in both root and shoot tissues at the seedling stage of *A. thaliana*. These cross-talk genes are presented with their protein identifiers and their molecular functions or biological processes in Tables below. References confirm that the identified cross-talk genes associated functions are indeed involved in the regulation or response to abiotic stresses in *A. thaliana*.

Table 4.8: A list of cross-talk genes under five abiotic stresses from *A. thaliana* root tissue at the seedling stage by PPI-integrated clustering approach.

At3g23240	**ATERF1(ERF1):** defense response, ethylene mediated signalling pathways, jasmonic acid mediated signalling pathway, sequence-specific DNA binding transcription factor activity. (Cheng *et al.*, 2013)
At2g25080	**Glutathione peroxidase: GPX1:** glutathione peroxidase activity, response to oxidative stress. (Glombitza *et al.*, 2004; Sugimoto and Sakamoto,1997; Rodriguez Malia *et al.*, 2003)
At3g08720	***A. thaliana* protein kinase 19:** abscisic acid mediate signal pathway, intracellular cellular signal transduction, protein kinase activity, protein phosphorylation, protein tyrosine kinase activity, response to chitin, response to cold, response to ethylene stimulus, response to heat, response to salt stress, transferase activity (Chinnusamy *et al.*, 2004; Abwao, 2012; Yin *et al.*, 2013).
At1g05100	**MAPKKK18:** protein kinase activity, transferase activity.(Yin *et al.*, 2013)
At4g25380	***A. thaliana* stress-associated protein 10:** response to cold, response to heat, response to salt stress, response to high light intensity, response to hydrogen peroxide, response to manganese ion, response to metal ion, response to nickel cation, response to zinc ion.
At1g08920	**ESL1:** carbohydrate transmembrane transporter activity, response to abscisic acid stimulus, response to salt stress, response to water deprivation.

At2g30490	**cytochrome P450 family gene (CYP73A5): oxidoreductase activity.** (Glombitza *et al.*, 2004)
At5g04250	**Cysteine proteinases superfamily protein:** cysteine-type peptidase activity, heat acclimation, response to fungus, response to jasmonic acid stimulus, response to wounding.
At4g04720	**ATCPK21:** calmodulin-dependent protein kinase activity, protein serine/threonine kinase activity, transferase activity. (Franz *et al.*, 2010; Abwao, 2012)
At5g46710	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At1g78070	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At5g24600	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At5g10695	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At5g50360	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)

Table 4.9: A list of cross-talk genes under five abiotic stresses from *A. thaliana* shoot tissue at the seedling stage by PPI-integrated clustering approach.

At1g80110	**ATPP2-B11:** carbohydrate binding. (Glombitza *et al.*, 2004)
At2g30490	**cytochrome P450 family gene (CYP73A5): oxidoreductase activity.** (Glombitza *et al.*, 2004)
At1g01470	**LEA14:** defense response to fungus, response to abscisic acid stimulus, response to cold, response to high light intensity, response to desiccation, response to water deprivation, response to wounding.
At3g05640	Protein phosphatase 2C family protein: catalytic activity, response to hyperosmotic salinity, protein serine/threonine phosphatase activity, response to abscisic acid, response to cold, response to water deprivation.

At3g23240	**ATERF1(ERF1):** defense response, ethylene mediated signalling pathways, jasmonic acid mediated signalling pathway, sequence-specific DNA binding transcription factor activity. (Cheng *et al.*, 2013)
At2g25080	**Glutathione peroxidase: GPX1:** glutathione peroxidase activity, response to oxidative stress. (Glombitza *et al.*, 2004; Sugimoto and Sakamoto,1997; Rodriguez Malia *et al.*, 2003)
At5g42050	**DCD(development and cell death):** MAPK cascade, defense response to fungus, response to hypersensitive, response to cold. (Chinnusamy *et al.*, 2004)
At3g05660	**ATRLP33**: MAPK cascade, kinase activity, negative regulation of defense response, response to heat, response to high light intensity, response to hydrogen peroxide, signal transduction. (Chinnusamy *et al.*, 2004; Sappl *et al.*, 2009; Yin *et al.*, 2013)
At2g47180	**ATGOLS1:** galactosyltransferase activity, response to abscisic acid, response to cold, response to heat, response to high light intensity, response to hydrogen peroxide, response to oxidative stress, response to salt stress, response to water deprivation, transferase activity. (Taji *et al.*, 2002; Sappl *et al.*, 2009; Abwao, 2012)
At2g47770	**ATTSPO:** response to abscisic acid stimulus, response to osmotic stress, response to salt stress. (Tuteja and Sopory, 2008)
At4g04720	**ATCPK21:** calmodulin-dependent protein kinase activity, protein serine/threonine kinase activity, transferase activity. (Franz *et al.*, 2010; Yin *et al.*, 2013; Abwao, 2012)
At1g67360	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At3g11420	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At1g67920	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)
At4g11220	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)

At1g07500	unannotated genes with unknown molecular functions and biological processes (with mutant lines available)

Conclusively, results shown in the above table illustrate that the PPI clustering method developed can not only effectively perform cluster analysis on short gene expression profiles, but can also produce results with significant biological meaning involved in general abiotic stress defense, known as cross-talk genes. These identified cross-talk genes were subjected to further validation to confirm their involvement in the key regulation of abiotic stress. Among the identified abiotic-stress tolerant genes in both the root and shoot tissues, a number of candidate genes are novel genes without annotation information. There are mutant lines available for all these novel genes according to the search of gene mutants in TAIR. Plant seeds can be ordered for those mutant lines in order to study the significance of identified novel genes for plant growth under abiotic stress conditions. Therefore, the list of identified cross-talk genes among all five abiotic stress conditions have provided very useful information for biologists to study gene functions and to identify genes associated with key regulation in abiotic stress response.

4.4 Conclusion

In this chapter, the PPI-integrated clustering method was presented to discover biological patterns from short time-series gene expression data integrated with PPI data. The motivation of integrating PPI data is to improve clustering effectiveness from only short time-series gene expression data.

To demonstrate the performance of my proposed method for clustering short time-series gene expression data, my method has employed on ten sets of publicly available gene expression data. Gene samples were taken from root and shoot tissues in *A. thaliana* that were treated under various abiotic stresses during the seedling stage. For applied datasets, stress-responsive clusters were evaluated using GO enrichment analysis for clustering effectiveness. Clustering results from applied 10 datasets were compared with results from STEM

algorithm for method performance. It is shown that my proposed method has better clustering effectiveness, in terms of the identification of directly stress-responsive genes and indirectly stress-induced genes in all stress-responsive clusters. Additionally, the identification of cross-talk genes across multiple stress-responsive clusters is significant for the maintenance of basic cellular functions involved in stress defense. In summary, the novel gene-clustering approach described in this Chapter can not only improve gene clustering effectiveness, but also contribute to functional prediction of identified candidate genes which are associated with abiotic-stress defense functions.

Chapter 5

Conclusions and Future Work

The short time-course microarray gene expression profiles present unique clustering challenges due to large amount of genes in the microarray dataset versus small numbers of time points in each gene variable. In this thesis, two novel clustering methods are developed, named network-based clustering and PPI-integrated clustering, which are ideal for clustering short gene expression profiles. The network-based clustering method is to generate gene co-expression network using CMI, which measures the non-linear relationship between genes. The similarity function of CMI considers both direct and indirect relationships between genes in the presence of other genes, which covers the defect of short gene expression profiles, by making use of more information in the dataset for calculating gene relationships. Therefore, the generated gene co-expression network can more comprehensively represent gene relationships than those that are calculated using linear similarity functions. The PPI-integrated clustering method is designed for clustering short gene expression profiles into multiple functional clusters by integrating PPI data. It considers gene expression profiles between each consecutive time point, and integrates PPI data for the purpose of result refinement. Integrating PPI data is the major advantage of this proposed method in comparison to other clustering algorithms that have been used for this purpose. Without PPI data refinement, many clustered patterns might give false positive results as we are missing the critical result correction step. Both the network-based approach and the PPI-integrated approach can definitely improve the issue of gene expression profiles being too short to produce effective results.

The stress-responsive clusters identified by network-based and PPI-integrated clustering methods reveal better method performance in comparison to ClusterONE and STEM, respectively, in terms of: (1) returning a relatively lower target GO *p-value*, (2) identifying higher

percentage of total stress-related genes and un-annotated genes in the stress-responsive clusters, and (3) identifying a number of commonly expressed stress-responsive genes across five abiotic stress conditions in both the root and shoot tissues of *A. thaliana*, so called cross-talk genes.

The overlapping scores (ω) of the stress-responsive clusters produced by the network-based and PPI-integrated clustering methods were calculated using Equation 5.1, as shown in Table 5.1 below. The overlapping scores of clusters produced from the two proposed methods calculate the amount of common genes between clusters generated from both methods for each dataset. As majority of the overlapping scores are above 0.5, results indicate that large amount of genes in the obtained clusters by two proposed methods are common.

$$\omega(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{5.1}$$

Table 5.1: Overlap scores of stress-responsive clusters produced by network-based and PPI-integrated clustering methods of ten datasets.

	Cold	**Drought**	**Heat**	**Salinity**	**Osmotic stress**
Root	0.72	0.61	0.67	0.72	0.58
Shoot	0.67	0.57	0.55	0.68	0.49

The proposed methods can apply for the prediction of gene functions and regulations, as well as the subtypes of cells. For the prediction of gene functions, methods can help study functions of many genes for which annotation information has not been previously available in the cluster. It is assumed that co-expressed genes in the same cluster are likely to be involved in the same cellular process for a particular function, and that the strong correlation of gene expression profile patterns between these co-expressed genes also indicate the gene co-regulatory relationships. To be more specific, an effective clustering method is also applicable for the genome-wide cluster analysis of plant species for studying mechanisms of plant defense responses. Results are extremely helpful in understanding the defense system

of plants in response to change of growth condition. Thus, to control the expression level of stress-responsive genes for better stress tolerance, in order to increase the productivity of staples such as canola, rice, corn and so on.

For the proposed methods, their ability to effectively cluster functionally related genes also helps to identify the gene regulatory motifs and cis-regulatory elements specifically to the gene cluster, which can yield hypotheses regarding the gene regulatory mechanism (Brazma and Vilo, 2000; Tefferi *et al.*, 2002). Therefore, gene cluster analysis can not only apply to datasets related to plants for stress response study, but can also apply to datasets related to human for disease detection, such as for the prediction of human-disease related gene clusters. Additionally, the identifications of gene regulatory motif and cis-regulatory elements in cluster are also extremely useful in drug design, in order to inhibit the transcription of disease-related genes.

Finally, clustering approach is able to identify new cell classes, such as new cancer classes through the clustering approach of cancer cells, known as the identification of subtypes of cells (Golub *et al.*, 1999). The clustering approach for class discovery can automatically separate distinct class of cancer cells without previous knowledge of these classes, such as to distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukaemia (ALL) in the DNA microarray gene expression data generated from a human acute leukemia sample (Golub *et al.*, 1999).

The idea of identifying cross-talk genes can be further applied to the identification of cross-talk genes across multiple species under a particular stress condition. This can help to understand the evolutionary distance between species, as the latest diverged species should have more functional genes in common (cross-talk genes) to activate the same defense system in species that are distantly related. Furthermore, the identification of cross-talk genes can be also applied to human disease treatment, such as cancer and diabetes. To identify the disease-related cross-talk genes among different development stages of the disease, different genders or different ages of patient population can efficiently deliver ideas for drug designing,

in order to develop drugs that can target various stages of the disease, different genders or ages of the patient populations.

As suggested below, there are a number of possible future directions based on the methods developed in this thesis.

First: because my network-based clustering method requires *a priori* information for seed selection, the method can be modified by *not* using *a priori* known stress-responsive genes as the source gene group for seed selection. Instead, the seed selection method from the original ClusterONE algorithm can be retained by selecting the gene with the highest degree of connectivity among the entire gene network. After that, instead of growing a cluster from the selected seed according to calculation of cluster cohesiveness, other objective functions could be used to grow a cluster, because it is possible to assume that clusters in networks are not structures with high internal connectivity, but low connectivity to the rest of network. One colleague (Bolin Chen) had used the PPI network to identify protein complexes and found that the protein complexes, rather than occurring as dense sub-structures, but appeared as star-like structures, comprising multiple dense structures with high intra-structure densities, but low inter-structure densities. Therefore, the same idea also can be applied to the network-based clustering method, in order to identify clusters from the generated gene co-expression network.

Second: future work could modify the path consistency algorithm to more efficiently generate the gene co-expression network. This can be done by obtaining a higher-order network step-by-step starting from the zero-order network, the algorithm could be modified to obtain the network in one step by predefining a network order. This would facilitate the generation of a higher-order network with fewer edges being deleted, thus can ensure that the generated network is not too sparse for later cluster analysis.

Third: more data types such as PPI data can be integrated for generating the gene co-expression network in the network-based clustering method. Such PPI data could either

be combined with microarray data prior to cluster analysis as cluster analysis input, or be integrated after cluster analysis for result refinement.

Fourth: more study could be done on the un-annotated genes in cluster for prediction of gene functions, and on cross-talk genes for validating their involvement in the key regulation of abiotic stress response.

Fifth: Other than *A. thaliana*, datasets of other species can be subjected to similar analysis using the proposed novel methods, such as yeast (*Saccharomyces cerevisiae*), *Caenorhabditis elegans*, *Mus musculus*, *Escherichia coli* and etc.

References

Abdi, H. (2007). Bonferroni and sidak corrections for multiple comparisons. In N.J. Salkind (ed.). *Encyclopedia of Measurement and Statistics*, Thousand Oaks, CA, Sage.

Abwao, S. I. (2012). Translational control of abiotic stress responses in Arabidopsis thaliana. PhD thesis, University of Glasgow.

Al-Shehbaz, I. A. and O'Kane Jr, S. L. (2002). Taxonomy and phylogeny of arabidopsis (brassicaceae). *The Arabidopsis Book*, 1-22.

Altay, G. and Emmert, S. F. (2010). Revealing differences in gene network inference algorithms on the network level by ensemble methods. *Bioinformatics*, **26**, 1738-1744.

Andre, S. R., Stuart, A. K., Jason, L. P., Bjorn, S. and Joshua, S. (2008). Mutual Information in Random Boolean models of regulatory networks. *Physical Review*, **77**(1).

Andrew, D. K. (2004). Graph clustering with restricted neighbourhood search (Master's thesis). Retrieved from Graduate Department of Computer Science University of Toronto.

Arai, K. and Barakbah, A. R. (2007). Hierarchical k means an algorithm for centroids initialization for k means. *Rep. Fac. Sci. Engrg. Saga Univ.*, **36**(1), 25-31.

Arenkov, P., Kukhtin, A., Gemmell, A., Voloshchuk, S., Chupeeva, V. and Mirzabekov, A. (2000). Protein microchips: use for immunoassay and enzymatic reactions. *Anal. Biochem.*, **278**, 123131.

Ashin, C. (2011). Cheers. *The second round album*. B'IN Music International CO., LTD.

Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R. and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet*, **37**(4), 382-390.

Boja, C. (2011). Clusters models, factors and characteristics. *International Journal of Economic Practices and Theories*, **1**(1).

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M. and Sherlock, G. (2004). GO TermFinder open source software for accessing gene ontology information and finding significantly enriched gene ontology term associated with a list of genes. *Bioinformatics*, **20**(18), 3710-3715.

Brazma, A., Hingamp, P., Quackenush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C. P., Klm, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)- toward standards for microarray data. *Nature Genetics*, **29**, 365-371.

Brazma, A. and Vilo, J. (2000). Minireview: Gene Expression Data Analysis. *Federation of European Biochemical Soc.*, **480**,17-24.

Buchanan, B. B., Gruissem, W. and Jones, R. L. (2002). Biochemistry and molecular biology of plants. *American society of plant physiologist*.

Bush, D.S. (1995). Calcium regulation in plant cells and its role in signalling. *Annu Rev Plant Physiol Plant Mol Biol*, **46**, 95122.

Butte, A., Tamayo, P., Slonim, D., Golub, T. and Kohane, I. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc Natl Acad Sci U S A*, **97**,12182-12186.

Cadeiras, M., Bayern, M. V., Sinha, A., Shahzad, K., Lim, W. K., Grenett, H., Tabak, E., Klingler, T., Califano, A. and Deng, M. C. (2010). Drawing networks of rejection - a systems biological approach to the identification of candidate genes in heart transplantation. *J Cell Mol Med*, **15**(4), 949-956.

Cheng, M. C., Liao, P. M., Kuo, W. W. and Lin, T. P. (2013). The arabidopsis ETHYLENE RESPONSE FACTOR1 regulates abiotic stress-responsive gene expression by binding to different cis-acting elements in response to different stress signals. *Plant Physiol.*, **162**(3), 1566-1582.

Chinnusamy, V., Schumaker, K. and Zhu, J. K. (2004). Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *J. Exp. Bot.*, **55**(395), 225-236.

Clarke, E. L., Loguercio, S., Good, B. M. and Su, A. I. (2013). A task-based approach for gene ontology evaluation. *J Biomed Semantics*, **4**(supple 1), S4.

Colcombet, J. and Hirt, H. (2008). Arabidopsis MAPKs a complex signaling network involved in multiple biological processes. *Biochem. J.*, **413**, 217-226.

Dalma-Weiszhausz, W. W., Warrington, J., Tanimoto, E. Y. and Miyada, C. G. (2006). The affymetrix Genechip platform: an overview. *Methods Enzymol.*, **410**, 3-28.

Dalman, M. R., Deeter, A., Nimishakavi, G., and Duan, Z. H. (2012). Fold change and *p-value* cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*, **13**(S11).

Dash, S., Hemert, J. V., Hong, L., Wise, R. P. and Dickerson, J. A. (2012). PLEXdb gene

expression resources for plants and plant pathogens. *Nucleic Acids Res.*, **40**(D1), 1194-1201.

Daub, C., Steuer, R., Selbig, J. and Kloska, S. (2004). Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, **5**, 118.

Defays, D. (1977). An efficient algorithm for a complete link method. *The Computer Journal (British Computer Society)*, **20**(4), 364366.

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene expression omnibus NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**(1), 207-210.

Eisena, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**(25), 14863 -14868.

Erik, L.C., Benjamin, M. G. and Andrew, I. S. (2013). A task-based approach for gene ontology evaluation. *Journal of Biomedical Semantics*, **4**(Suppl 1), S4.

Ernst, J., Nau, G. J. and Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, **21**, i159-i168.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases. Retrieved 17 December 2008.

Fraley, C., and Raftery, A. E. (1998). Technical report no. 329. Department of Statistics University of Washington Box 354322.

Franz, S., Ehlert, B., Liese, A., Kurth, J., Cazale, A. C. and Romeis, T. (2010). Calcium-dependent protein kinase CPK21 functions in abiotic stress response in *Arabidopsis thaliana*.

Mol. Plant, 1-14.

Glombitza, S., Dubuis, P. H., Thulke, O., Welzl, G., Bovet, L., Gotz, M., Affenzeller, M., Geisi, B., Hehn, A., Asnaghi, C., Ernst, D., Seidlitz, H. K., Gundlach, H., Mayer, K. H., Martinoia, E., Reichhart, D. W., Mauch, F. and Schaffner, A. R. (2004). Crosstalk and differential response to abiotic and biotic stressors reflected at the transcriptional level of effector genes from secondary metabolism. *Plant Molecular Biology*, **54**, 817-835.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Celler, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **288**(5439), 531-537.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society*, **28**(1), 100108.

Heinrich, K. W., Wolfer, J., Hong, D., LeBlanc, M. and Sussman, M. R. (2012) DNA millichips as a low-cost platform for gene expression analysis. *Plant Physiol.*, **159**(2), 548-557.

Hrmova, M., Lopato, S. (2013). Enhancing abiotic stress tolerance in plants by modulating properties of stress responsive transcription factors. In: Advances in Genomics of Plant Genetic Resources (Tuberosa R, Graner A, Frison E, eds.). *Springer Verlag.* In Press.

Huang, G. T., Ma, S. L., Bai, L. P., Zhang, L., Ma, H., Jia, P., Liu, J., Zhong, M. and Guo, Z. F. (2012). Signal transduction during cold, salt, and drought stresses in plants. *Mol Biol Rep.*, **39**(2), 969-987.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860921.

Joe, H. W., Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236-244.

Kaplan, F., Kopka, J., Haskell, D. W., Zhao, W., Schiller, K. C., Gatzke, N., Sung, D. Y. and Guy, C. L. (2004). Exploring the temperature stress metabolome of arabidopsis. *Plant Physiology*, **136**(4), 4159-4168.

Kavitha, V. and Punithavalli, M. (2010). Clustering time series data stream - a literature survey. *International Journal of Computer Science and Information Security*, **8**(1), 289-294.

Keiichi, M., Yukiko, U. Y., Takuhiro, Y., Tetsuya, S. and Kazuo, S. (2011). Global landscape of a co-expressed gene new work in barley and its application to gene discovery in triticeae crops. *Plant and Cell Physiology*, **52**(5), 785-803.

Kirk, P., Griffin, J. E., Ghahramani, Z. and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**(24), 3290-3297.

Knight, H. and Knight, M. R. (2001). Abiotic stress signalling pathways: specificity and cross-talk. *TRENDS in Plant Science*, **6**(6).

Kriegel, H. P., Kroger, P., Sander, J. and Zimek, A. (2011). Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, **1**(3), 231240.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Leao, D. Jr. *et al.* (2004). Regular conditional probability, disintegration of probability and radon spaces. *Proyecciones*, **23**(1), 15-29.

Lee. B. H., Henderson, D. A. and Zhu, J. K. (2005). The Arabidopsis cold-responsive tran-

scriptome and its regulation by ICE1. *The Plant Cell*, **16**, 3155-3175.

Liu, F., Kuo, W. P., Jenssen, T. K. and Hovig, E. (2012). Performance comparison of multiple microarray platforms for gene expression profiling. *Methods Mol Biol*, **802**, 141-155.

Liu, G., Wong, L. and Chua, H. N. (2009). Complex discovery from weighted PPI networks. *Bioinformatics*, **25**(15), 1891-1897.

Mann, M., Hendrickson, R. C. and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem.*, **70**, 437-473.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**(Suppl 1), S7.

Meyer, P., Lafitte, F. and Bontempi, G. (2008). minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information. *BMC Bioinformatics*, **9**, 461.

Michnick, S. W., Ear, P. H., Landry, C., Malleshaiah, M. K. and Messier, V. (2011). Protein-fragment complementation assays for large-scale analysis, functional dissection and dynamic studies of protein-protein interactions in living cells. *Methods Mol Biol.*, **756**, 395-425.

Nepusz, T., Yu, H. and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, **9**, 471-472.

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S. and Brazma, A. (2003). ArrayExpress a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*,

31(1), 68-71.

Pele, O., Taskar, B., Globerson, A. and Werman, M. (2013). The Pairwise Piecewise-Linear Embedding for Efficient Non-Linear Classification. *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013*. JMLR: W and CP volume 28.

Priness, I., Maimon, O. and Ben-Gal, I. (2007). Evaluation of gene-expression clustering via mutual information distance measure. *BMC Bioinformatics*, **8**, 111.

Rajaraman, A. (n.d.). Citing website. More data beats better algorithms. Retrieved March 24, 2008 from http://anand.typepad.com/datawocky/2008/03/more-data-usual.html.

Ricardo, Z. N. V. and Tie, K. (2005). HTself: self-self based statistical test for low replication microarray studies. *DNA Res.*, **12**(3), 211-214.

Rodriguez Milla, M. A., Maurer, A., Rodriguez Huete, A. and Gustafson, J. P. (2003). Glutathione peroxidase genes in Arabidopsis are ubiquitous and regulated by abiotic stresses through diverse signaling pathways. *Plant J.*, **36**, 602615.

Sappl, P. G., Carroll, A. J., Clifton, R., Lister, R., Whelan, J., Harvey Millar, A. and Singh K. B. (2009). The Arabidopsis glutathione transferase gene family displays complex stress regulation and co-silencing multiple genes results in altered metabolic sensitivity to oxidative stress. *Plant J.*, **58**(1), 53-68.

Seki, M., Narusaka, M., Abe, H., Kasuga, M., Yamaguchi-Shinozaki, K., Carninci, P., Hayashizaki, Y. and Shinozaki, K. (2001). Monitoring the expression pattern of 1,300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. *Plant Cell*, **13**, 6172.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (2007). Gene network involved in drought stress response and tolerance. *Journal of Experimental Botany*, **58**(2), 221-227.

Shinozaki, K. and Yamaguchi-Shinozaki, K. (1996). Molecular response to drought and cold stress. *Current Opinion in Biotechnology*, **7**, 161-167.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society)*, **16**(1), 30-34.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationship. *Univ. Kans. Sci. Bull.*, **28**, 1409-1438.

Stijn, V. D. (2000) Performance criteria for graph clustering and Markov cluster experiments. Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.

Stuart, J. M., Segal, E., Koller, D. and Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*, **302**(5643), 249-255.

Sugimoto, M. and Sakamoto, W. (1997). Putative phospholipid hydroperoxide glutathione peroxidase gene from *Arabidopsis thaliana* induced by oxidative stress. *Genes Genet. Syst.*, **72**, 311316.

Sun, P. G., Gao, L. and Han, S. (2011). Prediction of human disease-related gene clusters by clustering analysis. *Int J Biol Sci.*, **7**(1), 61-73.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P. and Huala, E. (2008). The arabidopsis information resource (TAIR) gene structure and function annotation. *Nucleic Acids Res.*, **36**(Database issue), D1009-1014.

Taji, T., Ohsumi, C., Luchi, S., Seki, M., Kasuga, M., Kobayashi, M., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2002). Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J.*, **29**(4), 417-426.

Tavazoie, S., Hughes, D., Campbell, M. J., Cho, R. J. and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genet*, **22**(3), 281285.

Tefferi, A., Bolander, E., Ansell, M., Wieben, D. and Spelsberg, C. (2002). Primer on Medical Genomics Part III: Microarray Experiments and Data Analysis. *Mayo Clinic Proc.*, **77**, 927-940.

The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. *Nucleic Acids Res.*, **36**(Database issue), D440-D444.

Trewavas, A. (1999). *Le calcium, cest la vie*: Calcium makes waves. *Plant Physiol*, **20**, 1-6.

Tu, Y., Stolovitzky, G. and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *PNAS*, **99**(22), 14031-14036.

Tuteja, N. and Sopory, S. K. (2008). Chemical signaling under abiotic stress environment in plants. *Plant Sgnal Behav.*, **3**(8), 525-536.

Walhout, A. J. and Vidal, M. (2001). High throughput yeast two hybrid assays for large scale protein interaction mapping. *Methods.*, **24**(3), 297-306.

White, J. R., Navlakha, S., Nagarajan, N., Ghodsi, M. R., Kingsford, C. and Pop, M. (2010). Alignment and clustering of phylogenetic markets implication for microbial diversity studies. *BMC Bioinformatics*, **11**(152).

Yeung, K. Y. and Ruzzo, W. L. (2001). An empirical study on principle component analysis for clustering gene expression data. *Bioinformatics*, **17**(9), 763-774.

Yin, Z. J., Wang, J. J., Wang, D. L., Fan, W. L., Wang, S. and Ye, W. W. (2013). The MAPKKK Gene Family in Gossypium raimondii: Genome-Wide Identification, Classification and Expression Analysis. *Int. J. Mol. Sci.*, **14**(9), 18740-18757.

Zhang, B. and Horvath, S. (2005). General framework for weighted gene coexpression analysis. *Stat Appl Genet Mol Biol*, **4**,17.

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., Hao, J. K., Liu, Z. P. and Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, **28**(1), 98-104.

Zhou, X., Kao, M. and Wong, W. (2002). Transitive Functional Annotation By Shortest Path Analysis of Gene Expression Data. *Proc Natl Acad Sci U S A*, **99**(20),12783-12788.

Appendix

Program code for correcting profile genes using PPI data (Step two of multiple data integration algorithm):

```
% open input file that each row is a cluster, columns are genes in cluster
% (The number of columns can be different in each row)
% file is readed in Profile_LineEle cell

Profile_fid = fopen('file.txt','r');


Profile_tline = fgetl(Profile_fid);
disp(Profile_tline);

Profile_LineEle = cell(1);

j=0;

while ischar(Profile_tline)


    Profile_ID = regexp(Profile_tline, '\s+', 'split');

    [m n]=size(Profile_ID);
```

```
    j=j+1;

        for i=1:n

            Profile_LineEle{j,i} = lower(Profile_ID{i});

        end

    Profile_tline = fgetl(Profile_fid);
    disp(Profile_tline);

end
fclose(Profile_fid);

% open PPI file and read PPI file into cell
%(PPI file is the general 2 columns PPI file)

PPI_fid = fopen('file.txt','r');

PPI_tline = fgetl(PPI_fid);
disp(PPI_tline);

PPI_LineEle = cell(1);

a=0;

while ischar(PPI_tline)

    PPI_ID = regexp(PPI_tline, '\s+', 'split');
```

```
    a=a+1;

        for b=1:2

            PPI_LineEle{a,b} = lower(PPI_ID{b});

        end

    PPI_tline = fgetl(PPI_fid);
    disp(PPI_tline);

end

fclose(PPI_fid);

N1 = cell(1);
N2 = cell(1);
[c n] = size(Profile_LineEle);

Score_matrix = zeros(c,n);

% Assigning score for each protein using PPI data

for j = 1:c
    for i = 1:n

        if isempty(Profile_LineEle{j,i}) == 0
            N1 = Profile_LineEle{j,i};
        else
            break
```

```
end

matrix = zeros (c,1);

for p = 1:c
    for q = 1:n

        if isempty(Profile_LineEle{p,q}) == 0
            N2 = Profile_LineEle{p,q};
        else
            break
        end

        data = FindID(PPI_LineEle,N1,N2);

        if data > 0
                matrix(p,1) = matrix(p,1)+data;

        else

        end

    end
    matrix(p,1) = (matrix(p,1))/q;
end

[Ca, Cb] = find(matrix == max(matrix));
ja = find(Ca == j);

if isempty(ja) == 0
```

```
            Score_matrix(j,i) = 0;

        else
            [num idx] = max(matrix(:));

            Score_matrix(j,i) = idx(1);
        end

    end

end

Profile_LineEle_changed = cell(c,1);

% Moving gene/protein to its new cluster based on correction of PPI data.

for j = 1:c
    for i = 1:n

        if Score_matrix(j,i) == 0
            num = j;
        else
            num = Score_matrix(j,i);
        end

            a = 0;

            [NewRow, NewCol] = size(Profile_LineEle_changed);
```

```
            for k = 1:NewCol
                if isempty(Profile_LineEle_changed{num,k}) == 1
                    a = k;
                    break;
                else
                end
            end

            if a == 0
                a = NewCol + 1;
            end

            Profile_LineEle_changed{num,a} = Profile_LineEle{j,i};

    end

end
```

Function used in the main program code:

```
function data = FindID(PPI_LineEle,N1,N2)
if strcmp(N1,N2)
data = 0;
else
j = length(PPI_LineEle(:,1));

idn1 = strcmp(PPI_LineEle,N1);
idn2 = strcmp(PPI_LineEle,N2);

data = sum(idn1(:,1)&idn2(:,2)) + sum(idn1(:,2)&idn2(:,1));
end
```